Digital Discrimination in the Sharing Economy: Evidence, Policy, and Feature Analysis

Miroslav Tushev
DIGITAL DISCRIMINATION IN THE SHARING ECONOMY: EVIDENCE, POLICY, AND FEATURE ANALYSIS

A Dissertation

Submitted to the Graduate Faculty of
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in
Division of Computer Science and Engineering

by
Miroslav Tushev
Specialist Degree, The Russian Academy of National Economy and Public Administration, 2014
M.S., Louisiana State University, 2018
May 2022
This dissertation is dedicated to my loving wife, Alina, for her patience and continued support during my research and writing.
Acknowledgments

I would like to thank my committee members, Dr. Gerald Baumgartner and Dr. Golden Richard III, for their time and support to this work. I would also like to thank Dr. Courtney Szocs, my dean representative, for agreeing to serve on my committee. A special thanks goes to my Academic Advisor, Dr. Nash Mahmoud, for his enormous support and mentorship during my Ph.D. Finally, I would like to acknowledge National Science Foundation for sponsoring this work.
# Table of Contents

Acknowledgments ................................................................. iii  

Abstract ............................................................................... v  

Chapter 1. Introduction .......................................................... 1  

Chapter 2. Literature Review ................................................... 5  
2.1. Introduction ................................................................. 5  
2.2. Method and Quantitative Analysis ..................................... 6  
2.3. Qualitative Analysis of Evidence ...................................... 12  
2.4. Discussion and Impact .................................................... 23  
2.5. Limitations and Threats to Validity ................................. 25  
2.6. Conclusion ................................................................. 26  

Chapter 3. User Perspective ...................................................... 27  
3.1. Introduction ................................................................. 27  
3.2. Data Scraping and Analysis ............................................. 27  
3.3. Results and Conclusion .................................................. 31  
3.4. Threats to Validity ........................................................ 32  

Chapter 4. Policies ................................................................. 34  
4.1. Introduction ................................................................. 34  
4.2. Motivation and Research Questions ................................. 35  
4.3. Data Collection ........................................................... 38  
4.4. Quality Assessment of NDPs ......................................... 44  
4.5. Results and Analysis ..................................................... 48  
4.6. Discussion and Impact ................................................... 57  
4.7. Conclusions ............................................................... 60  

Chapter 5. Framework ........................................................... 62  
5.1. Introduction ............................................................... 62  
5.2. Modeling Framework ................................................... 64  
5.3. Model Evaluation ......................................................... 72  
5.4. Extrinsic Evaluation ....................................................... 79  
5.5. Discussion and Impact .................................................. 84  
5.6. Limitations and Threats to Validity ................................. 86  
5.7. Conclusion and Future Work ......................................... 87  

Chapter 6. Conclusions ........................................................... 88  

Works Cited ........................................................................... 90  

Vita ......................................................................................... 102
Abstract

Applications (apps) of the Digital Sharing Economy (DSE), such as Uber, Airbnb, and TaskRabbit, have become a main facilitator of economic growth and shared prosperity in modern-day societies. However, recent research has revealed that the participation of minority groups in DSE activities is often hindered by different forms of bias and discrimination. Evidence of such behavior has been documented across almost all domains of DSE, including ridesharing, lodging, and freelancing. However, little is known about the underlying design decisions of DSE systems which allow certain demographics of the market to gain unfair advantage over others. To bridge this knowledge gap, in this dissertation, we investigate the problem of digital discrimination from a software engineering point of view. To develop an in-depth understanding of the problem, we first synthesize existing evidence on digital discrimination from interdisciplinary literature. We then analyze online user feedback, available on social media channels, to assess end-users’ awareness of discrimination issues affecting their DSE apps. We then introduce a novel protocol for drafting and evaluating nondiscrimination policies (NDPs) in the DSE market. Our objective is to assist DSE developers with drafting high quality and less ambiguous NDPs. Finally, we propose and evaluate a modeling framework for representing discrimination concerns affecting popular DSE apps along with their relations (synergies and tradeoffs) to other system features and user goals. Our objective is to visualize such complex domain knowledge using formal notations that software developers can easily understand, communicate, and utilize as an integral part of their app design process. The impact of the proposed research will extend to the entire population of DSE workers, targeting the deep racial and regional disparities in the DSE market and helping people in resource-constrained communities to overcome key barriers to participation and adaptation in one of the fastest growing software ecosystems in the world.
Chapter 1. Introduction

Over the past few years, the Digital Sharing Economy—also known as the sharing or gig economy—has become one of the most ubiquitous manifestations of mobile technology. Unlike conventional business models, applications of the Digital Sharing Economy (DSE) provide access to, rather than ownership of, underutilized assets and resources via Peer-to-Peer (P2P) coordination [1]. This on-demand, convenient, and ecologically sustainable form of resource consumption has attracted consumers and investors around the globe. As of today, there are thousands of the sharing economy platforms, enabling consumers to sell, rent, swap, lend, and borrow services and assets at unprecedented scales.

The unique form of direct business exchange that the sharing economy platforms have enabled has been linked to significant levels of economic growth, especially in communities at the lower end of the economic ladder, helping unemployed and partially employed individuals to generate income, increase reciprocity, and access resources that are unattainable otherwise [1, 2, 3, 4, 5]. However, recent research has exposed a serious discrimination problem affecting these platforms [6, 7, 8]. Discrimination, as a general term, refers to incidents where “members of a minority group (women, Blacks, Muslims, immigrants, etc.) are treated differentially (less favorably) than members of a majority group with otherwise identical characteristics in similar circumstances” [9]. In the context of the sharing economy, discrimination (also commonly known as digital discrimination) refers to a phenomenon where an online business transaction over a DSE platform is influenced (biased) by race, gender, age, or any other non-business related characteristic of service providers or consumers [10, 2, 6, 7, 8].

The problem of digital discrimination in online sharing economy markets has gained increasing attention in recent years. Numerous large-scale surveys, field studies, and data analysis papers have documented significant evidence on different patterns of discriminatory behavior across almost all domains of the sharing economy, including ridesharing (e.g., Uber), lodging (e.g., Airbnb), and freelancing (e.g., TaskRabbit). Such patterns include
discrimination based on ethnic background (racism), gender or sexual orientation (sexism), and physical disability (ableism) [6, 11, 7, 12, 13, 14, 15]. For instance, a recent study of ridesharing services found that Black riders using Uber waited on average 30% longer to be picked up [6]. Another study of P2P lodging services reported that non-Black Airbnb hosts were able to charge 12% more than Black hosts [7]. In the freelancing domain, a study of worker profiles on TaskRabbit revealed that the gender and race of workers were significantly correlated with their ratings [8]. This phenomenon is mainly facilitated by the P2P connection initiated between the sharing economy users, encouraging different forms of established bias (e.g., racism, sexism, and ableism) to transfer online [2, 6, 7, 8].

Limitations and Knowledge Gap. In traditional economy markets, discrimination is countered by imposing anti-discriminatory laws [16]. For instance, the U.S. Civil Rights Act of 1964 guarantees equal treatment of customers in public accommodations, such as hotels or rental property. However, in the cyberspace, discrimination takes a different form that is often difficult to detect and deter.

Existing research on digital discrimination often tackles the problem from a socio-economic and regulatory points of view [6, 7, 17, 18, 19, 20, 21]. In particular, researchers seek to prove and document discriminatory behavior in the DSE market as well as propose legislation to counter such behavior [2, 19, 22, 20, 16]. However, the research on the design aspects of DSE software which enable such a complex socio-technical phenomena to emerge online remains underdeveloped. In particular, we identify the following gaps in existing research:

- In software engineering research, digital discrimination is often tied to the problem of software fairness. Fairness research aims to propose methods for quantifying bias in software systems and to develop algorithmic solutions for fairness testing and preservation [23, 24, 25, 26, 27]. However, these solutions frequently ignore the design aspects of DSE platforms which allow long-standing issues of offline bias to flourish online [17, 19]. This emphasizes the need for a fundamental shift in the way DSE
developers think about their design. Specifically, discrimination concerns should be explicitly defined and considered as an integral part of the app design process.

- Existing research on digital discrimination relies on direct user surveys (questionnaires and interviews) and field studies to identify discrimination concerns among DSE users [6, 7, 8]. However, these data collection methods are extremely costly, and the sample size, or response rate, are often limited by factors such as the geographical area covered, number of subjects surveyed, and number of platforms studied. To overcome these limitations, other, more wide-spread channels of user feedback, should be exploited to examine the prevalence of discrimination concerns over these instant sources of data.

- Unlike software privacy policies, there is a lack of knowledge about how non-discrimination policies (NDPs) can be drafted and structured. This can be attributed to the fact that NDPs are non-code artifacts, thus, creating and evolving such artifacts often fall outside of developers’ expertise.

Outline and Contributions. To bridge the knowledge gap in existing research, this dissertation presents a first-of-its-kind analysis of the problem of digital discrimination from an end-user, software developer, and policy-maker perspectives. In particular, the main contributions (outline) of this dissertation can be described as follows:

- In Chapter 2, we systematically synthesize evidence from 58 interdisciplinary studies to identify the pervasive discrimination concerns affecting DSE platforms along with their triggering features and mitigation strategies. Our objective is to consolidate such interdisciplinary evidence from a software design point of view.

- In Chapter 3, we collect and qualitatively analyze a large-scale dataset of user feedback scraped from the Twitter feeds of eight popular DSE platforms. Our objective is to examine the distribution and types of bias reported in the online feedback of end-users of DSE platforms as well as assess their awareness of the problem.
• In **Chapter 4**, we introduce a systematic protocol for analyzing and evaluating the content of NDPs in the DSE market. Our analysis is conducted using a dataset of 108 DSE apps, sampled from a broad range of application domains. Our objective is to aid DSE app developers with drafting and evolving more comprehensive NDPs as well as help end-users of these apps to make more informed socio-economic decisions that can lead to optimized outcomes.

• In **Chapter 5**, we propose and empirically evaluate a conceptual framework for modeling discrimination concerns in the DSE market. The proposed framework utilizes domain modeling techniques to represent concerns of digital discrimination in the DSE market along with their relations to the functional features and user goals of DSE platforms. Our objective is to present such a complex socio-technical domain phenomena using simplified notations that software engineers can easily interpret, communicate, and use as an integral part of their app design process.

• In **Chapter 6**, we conclude the dissertation and discuss directions of future work.
Chapter 2. Literature Review

Applications of the sharing economy, such as Uber, Airbnb, and TaskRabbit, have become a main facilitator of economic growth and shared prosperity in modern-day societies. However, recent research has revealed that the participation of minority groups in the sharing economy activities is often hindered by different forms of bias and discrimination. Evidence of such behavior has been documented across almost all domains of the sharing economy, including ridesharing, lodging, and freelancing. However, little is known about the underlying design decisions of the sharing economy platforms which allow certain demographics of the market to gain unfair advantage over others. To bridge this knowledge gap, in this chapter, we systematically synthesize evidence from 58 interdisciplinary studies to identify the pervasive discrimination concerns affecting the sharing economy platforms along with their triggering features and mitigation strategies. Our objective is to consolidate such interdisciplinary evidence from a software design point of view. Our results show that existing evidence is mainly geared towards documenting and mitigating issues of racism and sexism affecting platforms of ridesharing, lodging, and freelancing. Our review further shows that discrimination concerns in the sharing economy market are commonly enabled by features of user profiles and commonly impact reputation systems.

2.1. Introduction

The sharing economy refers to a sustainable form of online business exchange that is built around sharing assets and resources rather than transferring their ownership [1]. Over the past decade, applications of the sharing economy, such as Uber, TaskRabbit, and Airbnb, have caused major disturbances in established classical markets, enabling people to exchange and monetize their underused (or idle) assets and skills at an unprecedented scale [28, 18, 2]. As of today, there are thousands of active sharing economy platforms, operating in a market sector that is projected to grow to close to 335 billion U.S. dollars by 2025 [29].
Existing research on digital discrimination often tackles the problem from socio-economic and regulatory points of view [6, 7, 17, 18, 19, 20, 21]. In particular, researchers seek to prove and document discriminatory behavior in the sharing economy market as well as propose legislation to counter such behavior [2, 19, 22, 20, 16]. However, the research on the design aspects of the sharing economy software which enable such a complex socio-technical phenomenon to emerge online remains underdeveloped. This can be partially attributed to the fact that existing evidence on digital discrimination is scattered across a broad range of interdisciplinary venues. Locating, interpreting, and synthesizing such evidence can be a very challenging task, especially in highly agile environments where the main focus is on solution deployment rather than problem research.

To bridge this knowledge gap, in this chapter, we conduct a first-of-its-kind effort to systematically consolidate a large body of interdisciplinary research on digital discrimination, a complex socio-technical problem that is currently affecting millions of users in one of the fastest growing software ecosystems in the world. Our objective is to facilitate evidence-based software design strategies [30] by helping the sharing economy developers to a) identify the main discrimination concerns in their domain of operation, b) understand how the interactions between their functional features and user goals can facilitate bias and differential treatment of the sharing economy users, and ultimately c) deliver the sharing economy solutions that can promote equality and mitigate bias by design.

The rest of this chapter is organized as follows. Section 2.2 describes our review protocol and presents a quantitative analysis of existing evidence. Section 2.3 qualitatively synthesizes available evidence on digital discrimination in the sharing economy literature. Section 2.4 discusses our main findings. Section 2.5 addresses threats to validity. Finally, Section 2.6 concludes the chapter and describes our future work.

2.2. Method and Quantitative Analysis

The research on digital discrimination in the sharing economy market aims to provide strong empirical evidence on the different patterns of bias affecting different sharing
economy platforms and suggest feature changes to enhance these platforms’ resilience to inequality. To locate such evidence, in this section, we conduct a systematic literature review (SLR) of this body of research. According to Kitchenham et al. [31], SLR as a research methodology consists of three main steps: planning, conducting, and reporting. Under the planning phase, the need for the review is justified, the review protocol is established, and the research questions are defined. During the conducting phase, the review protocol is put into action, including the identification of primary studies and categorizing and synthesizing existing evidence. Finally, under the reporting phase, the results are reported in a way that is tailored for the intended audience. In what follows, we describe our review protocol in greater detail.

2.2.1. Research Questions

It is essential to identify a set of research questions before taking on a review study. Research questions are necessary to identify the scope of studies (papers) to be included in the search process and to outline the objectives of the review. In this chapter, our research questions are:

- **RQ1**: *What types of discriminatory behavior do the sharing economy platforms exhibit?* Discrimination can take many forms; some are more prominent than others. Therefore, under this research question, we seek to determine the specific types of discriminatory behavior, or bias, that are common in the sharing economy market.

- **RQ2**: *What domains, or platforms, of sharing economy are affected the most by discrimination?* Sharing economy platforms extend over a broad range of application domains, from ridesharing, to lodging, and even dog walking (e.g. Wag!). Therefore, under this question, we seek to identify the application domains of the sharing economy that are commonly affected by discrimination.
• **RQ3:** *What are the main features and user goals that are related to discrimination in the sharing economy platforms?* This research question is concerned with synthesizing evidence on the underlying design decisions and feature-goal interactions of the sharing economy platforms that are responsible for enabling discriminatory behavior.

• **RQ4:** *Are there any suggested feature changes to counter digital discrimination?* Under this research question, we seek to locate evidence on any design strategies that have been suggested to counter, or mitigate, the different types of discrimination prevalent in the sharing economy market.

### 2.2.2. Identifying Primary Studies

To identify our set of primary studies, we start by formulating our search query. The most common term that is often used to refer to discriminatory behavior in the sharing economy is *digital discrimination*. To account for the variations and synonyms of *discrimination*, we refer to the Oxford English Dictionary. The following synonyms were included to our search query: *bias, prejudice, inequity, and bigotry*. In addition to these generic terms, we consider specific types of discrimination—in case a primary study referred to a specific type of discrimination and not the word *discrimination* or its synonyms. Table 2.1 lists the main acts of discrimination as described by the U.S. Equal Employment Opportunity Commission. These acts commonly appear in diversity and social justice literature [32]. Based on this list, we add the terms *racism, sexism, ableism, ageism, parental, classism, and religious*. Given that some of these types are more common than others, we also include variations for less popular types, such as *disability* and *accessibility* for ableism and *LGBT* for sexism. We further enhance our query with information about the popular domains and platforms of the sharing economy, including ridesharing (Uber and Lyft), lodging (Airbnb, Couchsurfing, and Vrbo), freelancing (TaskRabbit, Fiverr, and Upwork), and food delivery (DoorDash, Postmates, and UberEats). Note that these terms are chained using an (OR) command to avoid omitting any other, less popular, domains or platforms. Finally, to make
sure we are being specific to the domain of the sharing economy, we add the terms sharing economy, gig economy, and shared economy. In summary, our query can be described as follows:

\[
\text{((Digital discrimination) OR discrimination OR bigotry OR bias OR prejudice OR inequity OR racism OR sexism OR LGBT OR ableism OR ageism OR parental OR classism OR religious OR disability OR accessibility) AND (ridesharing OR Uber OR Lyft OR Lodging OR Airbnb OR Vrbo OR food delivery OR Doordash OR UberEats OR Postmates OR freelancing OR TaskRabbit OR Fiverr OR Upwork OR (sharing economy) OR (Shared Economy) OR (gig economy))}
\]

Our search was conducted over Google Scholar, the ACM Digital Library, IEEE Xplore, and Scopus. The results of the search were iteratively examined to add more terms to the query and explore more research venues. The process stopped when no more new primary studies were found [33]. In total, 84 papers were located using our iterative search process.

2.2.3. Inclusion and Exclusion Criteria

In SLRs, the inclusion and exclusion criteria are used as a basis for selecting primary studies. Such criteria should be determined beforehand during the planning phase. Our inclusion criteria in this chapter are:

- Books, papers, and technical reports.
- Studies that explicitly investigate design issues of digital discrimination in the sharing economy.
- Studies that are published in English.

We used the following exclusion criteria to exclude any studies that are irrelevant to our survey goals:
Table 2.1. Most common types of discrimination according to the U.S. Equal Employment and Opportunity Commission.

<table>
<thead>
<tr>
<th>Type</th>
<th>Discrimination against:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Racism</td>
<td>Ethnicity, color, or nationality.</td>
</tr>
<tr>
<td>Sexism</td>
<td>Gender or sexual orientation.</td>
</tr>
<tr>
<td>Ableism</td>
<td>Physical, sensory, or intellectual disability.</td>
</tr>
<tr>
<td>Parental</td>
<td>Parents with children or pregnant women.</td>
</tr>
<tr>
<td>Ageism</td>
<td>Older or younger people.</td>
</tr>
<tr>
<td>Religious</td>
<td>Perceived religion or a set of beliefs.</td>
</tr>
<tr>
<td>Classism</td>
<td>Particular social class.</td>
</tr>
</tbody>
</table>

- Short papers (less than 4 pages), editorials, summaries of keynote, tutorial papers, and grey literature.

- Duplicate reports of the same study. In case of duplication, the most recent version is selected.

To include and exclude papers, each paper was examined by each of the three authors individually. Specifically, each author read the title, abstract, and if necessary, the body of each of the 84 papers to determine their relevance to our survey. Each judge flagged each paper as Include (IN), Neutral (NU), or Exclude (EX). The paper was then included or excluded based on the protocol shown in Table 2.2. Cases of conflicts were resolved using majority voting. Applying our inclusion/exclusion criteria to our initial round of search resulted in 40 papers (48%). Our main observation during this process is that a large number of papers were specific to regulatory issues, or legislation to enforce equality in the sharing economy markets, with no discussion of aspects related to platform design. These papers were excluded.

To reduce the risk of omitting relevant studies, we also performed a lightweight backward-forward-snowballing on the included papers [34]. We basically inspected the studies cited by each of our included primary studies and the publications that subsequently cited the
study. In total, 18 more papers were identified, raising the number of our studies to 58 papers. We did not enforce a venue criterion on our primary studies, mainly because the problem itself is inherently interdisciplinary, thus, enforcing specific venues might lead to omitting important related work.

2.2.4. Quantitative Analysis

We start our review by performing basic quantitative analysis on our included studies. This involved each of the authors individually going through each study to determine the type of discrimination the paper tackles, the specific the sharing economy domain or platform being investigated in the paper, and the research methodology used. A discussion session was then held to consolidate our findings.

With regard to $RQ_1$ and $RQ_2$, our results show that discrimination problems are mainly investigated in the domains of ridesharing, lodging, and freelancing. In terms of platforms, Airbnb and Uber are the most investigated platforms (Fig. 2.1). This actually was expected given that discrimination concerns are more likely to manifest over such popular platforms as they tend to have significantly larger and more heterogeneous userbases in comparison to less popular platforms. Our results also show that racism and sexism are the most common types of discrimination investigated in the literature. Such studies started appearing early in the past decade, before taking off in 2015 (Fig. 2.2). A specific index of these papers is shown in Table 2.3. In terms of methodology, our review shows that the majority of primary studies on digital discrimination take the form of field stud-
Figure 2.1. Distribution of studies over the sharing economy platforms.

Figure 2.2. A bubble chart of the growth of the digital discrimination literature over time.

ies [35, 36, 6, 37] and large-scale surveys [38, 11]. Studies that rely on analyzing online platform data (user reviews or service listings) are also common [39, 40, 41, 14, 8].

2.3. Qualitative Analysis of Evidence

A major goal of our SLR is to synthesize evidence on the features or goals of the sharing economy platforms that have been proven to enable discriminatory behavior in the sharing economy market (RQ₃) as well as their mitigation strategies (RQ₄). A functional feature can be described as any observable behavior of the system that satisfies a specific stakeholder need, and a user goal, or a softgoal, can be defined as any abstract user objective that the system should achieve [65]. Unlike functional features, user (soft) goals do not have a clear-cut criterion for their satisfaction, however, they can be partially met, or
Table 2.3. Included papers by discrimination type.

<table>
<thead>
<tr>
<th>Type</th>
<th>Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Racism</td>
<td>[6, 7, 17, 20, 42, 43, 44, 45, 46, 36, 37, 47], [48, 15, 8, 41, 49, 50, 51, 35, 21, 52, 53, 54, 55]</td>
</tr>
<tr>
<td>Sexism</td>
<td>[8, 12, 56, 14, 40, 57, 50, 53, 51, 58, 35, 55, 59]</td>
</tr>
<tr>
<td>Ableism</td>
<td>[60, 61, 62, 63]</td>
</tr>
<tr>
<td>Classism</td>
<td>[56, 14, 41, 13, 64]</td>
</tr>
</tbody>
</table>

satisfied through functional features [66, 67, 68]. To extract such evidence, we utilize a grounded theory approach of open coding and memoing [69]. This process can be described as follows:

- Each member of the review team (three authors) examined the title, abstract, and body of the paper. The main goal is to extract evidence on RQ3 and RQ4.

- Categories of evidence were recorded as they emerged in the text. Reviewers used memoing to keep track of the reasoning behind their categorization.

- An axial coding session was then held to consolidate individual reviewers’ categorizations into more abstract categories.

- Generated categories were then iteratively revised until no more categories or evidence were found.

By the end of our analysis, two main categories of features (profile information and reputation systems) and four user goals (trust, safety, accessibility, and inclusion) have emerged. These categories are described next.

### 2.3.1. Profile Information

The sharing economy platforms use users’ personal (profile) information as a means to enable effective search for service providers and receivers as well as to reduce anonymity and
facilitate identification offline [39, 41]. However, our review revealed that user profiles were commonly associated with digital discrimination. Basically, service providers or receivers can decline or cancel a transaction based on certain physical traits, such as ethnicity, gender, or age, that can be inferred from profile pictures, user names, or location [49, 35]. In what follows, we review evidence related to patterns of digital discrimination enabled by the sharing economy user profile information as well as the main strategies to counter these patterns.

Evidence. In the literature, profile pictures have been mainly linked to racism and sexism. For instance, based on an empirical analysis of 395 Airbnb's listings, Ert et al. [39] found that more trustworthy-looking Airbnb hosts charged higher prices for similar apartments. In another study of 200 U.S. consumers, Su and Mattila [40] reported that female consumers were more likely to book an Airbnb property listed with female profile pictures. In a study of 1,020 Airbnb listings, Jaeger et al. [41] reported that photo-based impressions of hosts' attractiveness significantly influenced their rental prices. The authors also reported that Black hosts charged lower prices for their apartments compared to White hosts. In a more recent field study of 100,000 Airbnb profiles across 24 cities, 14 countries, and 3 continents, Jaeger and Sleegers [43] found that personal information about sellers, as inferred from their names and pictures, led to widespread discrimination against hosts from racial minorities. In fact, racial profiling based on pictures is not specific to African-Americans; other independent large-scale studies have reported significant photo-induced discrimination against Asian and Hispanic hosts on Airbnb [45, 46]. For instance, in a recent study of hiring biases in freelancing, Leung et al. [35] asked 206 subjects to make hiring decisions for a mathematically intensive task. Significant biases against Black workers and less attractive workers and preferences towards Asian workers and women workers were detected.

Similar to profile pictures, user names were also linked to sexism and racial profiling [50]. For instance, according Foong et al. [56], Upwork workers with a unisex or unidentifiable
name had a $2.26 higher mean bill rate than female users on average. Another study by Barzilay and Ben-David [14] showed that women’s average hourly rates on P2P freelancing platforms were about two-thirds of men’s rates. In a field experiment of 1,801 Airbnb hosts, Cui et al. [36] found that requests from Airbnb guests with Black-sounding names were 19.2% less likely to be accepted than those with White-sounding names. In another field study of 6,400 Airbnb requests, Edelman et al. [7] reported that guests with Black-sounding names were 16% less likely to be accommodated relative to identical guests with White-sounding names. Ge et al. [6] conducted a field study of 1,500 UberX and Lyft ride requests on controlled routes. The authors observed more frequent cancellations by Uber drivers against passengers with Black-sounding names. These findings seem to persist globally [52, 53, 37]. For instance, in a field experiment of 952 carpooling requests in Germany, Carol [53] observed that women with German names were least likely to experience discrimination, while men with Turkish names were the most likely to face discrimination. Another experiment of 1,599 Airbnb requests in Norway showed that guests would spend less money on an apartment when the host was “Abdi” from Somalia rather than “Martin” from Norway [37]. Names were also found responsible for discrimination against the LGBT community. For instance Ahuja and Lyons [12] analyzed Airbnb host responses to listings indicative of LGBT relationships (e.g., “My name is (male/female name) and my (boyfriend/girlfriend) and I are ..”). The results showed that hosts were more likely to not reply at all rather than replying “no” to male-male pairs inquiring about room availability.

Profile information was also found responsible for other, less popular, types of discrimination such as classism, ageism, parental, and ableism. For instance, Moody et al. [11] surveyed 1,100 of UberPOOL and Lyft riders. The results showed that White passengers who lived in predominantly White communities were more likely to discriminate against other passengers they perceived to belong to a lower social class. In another study of classism, Thebault et al. [13] surveyed workers on TaskRabbit from the Chicago metropolitan area. The authors found that requests from customers in the socioeconomically disadvan-
taged South Side area were less likely to be accepted. As an example of ageism and parental
discrimination, a survey of 192 Airbnb hosts by Karlsson et al. [38] found that hosts were
more likely to accept older people and women. The survey also found that couples with a
child in their profile pictures were disadvantaged. In a study of discrimination against peo-
ple with disabilities, a randomized field experiment of 3,847 lodging requests by Ameri et
al. [63] revealed that hosts were less likely to approve requests from travelers who declared
blindness, cerebral palsy, dwarfism, or spinal cord injury in their profiles than to approve
travelers without disabilities.

**Mitigation Strategies.** Our review has uncovered several design strategies that have
been proposed and evaluated in the literature to control for discrimination issues stemming
from profile information. These strategies can be described as follows:

- **Withholding Information.** The main solution to counter profile-induced discrimi-
  nation is to minimize visual or verbal cues of users, allowing transactions to happen
  in relative anonymity [19]. For example, Uber prevents drivers from learning the
  identity or destination of their clients until they accept a request. Withholding, or
delaying the exposure, to such information was found to have a significant positive
effect on minimizing discrimination [70]. For instance, Mohammed [47] evaluated
Airbnb’s policy of delaying the exposure of guests to hosts’ profile photos in four
U.S. cities. The results provided a clear evidence on the success of this redesign in
narrowing the racial booking gap in Airbnb.

- **Self-disclosure information.** While withholding pictures and names can miti-
gate discrimination, entirely concealing such information is expected to deteriorate
  safety [71]. To work around this dilemma, a recent study suggested that Airbnb hosts
who discussed self-disclosure topics in their profiles, such as their tastes in music and
food, work, or study, were often perceived as more trustworthy, thus were more likely
to be chosen as hosts [72, 44]. In other words, *non-deterministic* information helps
to alleviate racial profiling by enabling a more humane perception of users as well
as challenging stereotypes [45]. In general, converging evidence suggests a redesign of user profiles, where information indicative of ethnicity, religion, or gender are hidden until after the transaction is confirmed, while more self-disclosure information (e.g., socially rich pictures or emotionally dynamic text [73]) are provided to reduce uncertainty and signal trustworthiness [19, 72].

- **Asset-based profile pictures.** In their field study, Hannák et al. [8] reported that workers who did not use a profile picture at all received significantly smaller numbers of reviews. The impact of not using a picture at all was also studied by Tjaden et al. [54] who found that Arab/Turkish/Persian drivers without a profile picture were observed to be much more disadvantaged than drivers from the same ethnic group with a profile picture. However, according to Ert e al. [70], workers who used a picture that was related to their assets (rental place as in Fig. 2.3) or skills (advertisements for the worker’s task) but not a face picture did not experience such decline. These findings suggest that workers can use their profile pictures to emphasize their skill while obfuscating their true demographics but without negatively affecting their reputation [70].

- **Fully automated matching.** Another mitigation strategy of profile-induced discrimination is to fully automate the P2P matching process. For instance, Uber riders do not have the luxury to choose from a list of nearby drivers. In Airbnb, the *Instant Booking* feature enables a guest’s request to be automatically accepted without an explicit consent action from the host. Hosts who enroll in this option are rewarded with a better search placement and Superhost status. According to several studies, this design decision helps to minimize the chances of biased assessment as service providers and receivers do not engage in any negotiations beforehand [12, 51].

- **Cashless payments.** In a study of 1,704 Uber, Lyft, and taxi trips in Los Angeles, Brown [48] reported that cashless payments in P2P ridesharing services may counter
racial discrimination. In particular, drivers indicated that paying through the app eliminated fears of fare evasion, thus reduced proxy discrimination against Black male riders.

2.3.2. Reputation Systems

Reputation systems (ratings and reviews) are considered the de facto trust-building mechanisms in the sharing economy [42, 74, 59]. However, our review of existing evidence revealed that the current design of these systems can enable discrimination. In particular, the aggregated reputation scores often reinforce prior discrimination beliefs of the sharing economy users [55, 58, 75]. In what follows, we review evidence related to patterns of digital discrimination affecting reputation systems of the sharing economy platforms as well as the main strategies to counter these patterns.

Evidence. Hannák et al. [8] analyzed 13,500 worker profiles on TaskRabbit and Fiverr. They found that Black workers received worse ratings and fewer reviews than similarly qualified White workers. The authors also analyzed linguistic bias in textual reviews. They observed that reviews for workers perceived to be Black women included significantly fewer positive adjectives, while reviews for Black workers contained significantly more negative adjectives. These results were remarkably consistent after controlling for platforms and cities from which the data was collected. In a more recent study, Goel et al. [55] analyzed a dataset of 8,218 listings on Airbnb from New York City, including 5,716 listings from White hosts and 2,502 from non-White hosts. The results confirmed that the ethnicity of the host and the majority ethnicity of the neighborhood had a significant effect on ratings and prices.
Our review also showed that bias in ratings and reviews influenced minorities’ participation in the sharing economy. For instance, Teubner et al. [74] analyzed 15,198 Airbnb listings from 86 German cities. They found that reputation, quantified through higher ratings and higher number of reviews, actually translated into significant economic value, either by attracting more demand or by allowing hosts to set higher listing prices. Several explanations were proposed for this phenomenon. For instance, Hannák et al. [8] reported that bad reviews or ratings often led to lower search ranks in freelancing platforms. In their field experiment, Cui et al. [36] reported that positive reviews posted on Airbnb guests’ pages significantly reduced discrimination towards guests’ with Black-sounding names. In addition, in an experiment with 8,906 Airbnb users, Abrahao et al. [59] reported that having a decent reputation was enough to counteract homophily, or the tendency of people to prefer or seek others who are similar. Another study by Brown et al. [48] reported that rider ratings may reduce proxy discrimination by drivers as they can use star ratings to infer how safe or considerate a rider may be.

**Mitigation Strategies.** Similar to profile information, several design strategies have been proposed in the literature to control for bias affecting reputation systems. These strategies (functional measures) can be described as follows:

- **Mutual reviews.** To prevent biased reviews, Airbnb rolled out a design change to ensure that hosts and guests can see the reviews only after both parties have submitted their reviews. According to Airbnb, “*Both hosts and guests may worry that if they leave an honest review that includes praise and criticism, they might receive an unfairly critical review in response. To address this concern, reviews will be revealed to hosts and guests simultaneously.*” This change was evaluated by Ert et al. [70] through an independent field study. The results showed that hiding reviews until the other party submitted their reviews significantly reduced discriminatory charged text in reviews.
• **Structured reputation systems.** A suggested design strategy to mitigate bias in reviews is to eliminate free-text reviews altogether. Instead, feedback should be structured in a set of predefined fields where input categories, along with acceptable inputs for each category, are provided [64]. While this change does not entirely eliminate bias, it can at least limit subjective reviews. For example, in a field experiment of 952 entry-level workers from Upwork, Pallais [76] observed that providing more structured (objective) evaluations substantially limited sentiment in reviews and improved workers’ subsequent employment outcomes. This can be very critical for service providers as a study of 47,651 Airbnb listings and 1,014,134 reviews found that guests, especially female travelers, were likely to be influenced by the sentiment of reviews [57].

• **Explicit trust cues.** The controlled experiments conducted by Nødtvedt et al. [37] showed that racial discrimination disappeared in the presence of an explicit trust cue, giving an indication that textual reviews could be effectively replaced by performance badges (e.g., Superhost or Elite Tasker). Along the same lines, Tjaden et al. [54] conducted a field study of 16,624 real carpooling rides from Germany. The results showed that additional objective cues about users, such as measured experience, can decrease the magnitude of ethnic discrimination and act as trust signal for consumers.

• **Hiding older reviews.** Emerging evidence suggests that the sharing economy platforms could consider showing only the most recent reviews for each user, while hiding the rest along with the total number of reviews per user. According to Hannák et al. [8], this design decision can level the playing field for workers, while still providing timely and testimonial feedback. These results were confirmed by Qiu et al. [77] who found that hiding the number of reviews on a platform such as Airbnb helps to avoid systematically disadvantaging newer users, yet also ensure that biases displayed by users are kept in check.
• **Bias-free rating elicitation.** Goel et al. [55] implemented an incentive mechanism to elicit fair ratings from users. The method utilizes a peer-consistency mechanisms known as the Peer Truth Serum for crowdsourcing [78]. The authors provided significant proof that such reward mechanisms can encourage users to try the service of individuals belonging to disadvantaged social classes and at the same time elicit truthful ratings about the quality of service received.

• **Bias correction.** This feature involves adjusting individual worker’s ratings to compensate for measurable sources of bias. In particular, since biases do exist, and can be effectively quantified, their effect can be reversed by adjusting rating scores for minority individuals [8]. In their work on reputation systems biases, Goel et al. [55] used the covariance between the aggregated reputation scores and the ethnicity as a proxy to measure bias. Applying the proposed transformation on a dataset of Airbnb reviews showed that adjusting for sensitive attributes such as ethnicity removed their impact, while the impact of other relevant attributes remained significant.

### 2.3.3. User Goals

User goals in the sharing economy can range from economic growth to ecological sustainability to building up social capital [2]. Our qualitative analysis of existing literature has exposed four types of user goals that are explicitly related to digital discrimination. The distribution of these goals over our primary studies is shown in Table. 2.4. These goals are:

• **Inclusion.** Inclusion (participation or equality) can be considered the antidote of discrimination. All included studies are geared towards addressing this goal. Ultimately, users want to be able to engage in the sharing economy activities as service receivers or providers without being treated differently for reasons unrelated to the nature of the transaction.
• **Trust.** Trust refers to the willingness of a party to be vulnerable to the actions of another party [79]. Our review has revealed that trust is a key user goal in the sharing economy [80, 81, 70, 40, 57]. The concept itself stems from the fact that conducting business transactions with uncertified strangers involves inherent risk, therefore, providers and receivers at both ends of the P2P connection need to establish a certain level of mutual trust before a transaction can take place [82, 39]. In fact, almost all platforms in our domain provide a trust statement on their websites, listing all the measures taken to establish trust in the platform and its users, such as reviews and ratings.

• **Safety.** As mentioned earlier, the sharing economy is inherently risky. Therefore, safety is another major goal of the sharing economy users. For example, riders need to feel safe before they get into a stranger’s car and hosts need to trust that guests would not harm their families or destroy their property. In the digital discrimination literature, safety is often referenced indirectly [22, 83, 44]. For instance, while profile information is used to enforce user safety (establishing trust and enabling identification offline), such information can be a main enabler of discrimination. Overall, the relation between safety as a user goal and discrimination as a user concern and the nature of interaction between them is still unclear.

• **Accessibility.** Accessibility is another major user goal for people with disabilities as well as parents. Accessibility in our context refers to the accessibility of the service itself. Primary studies tackling ableism emphasize accessibility as a main user goal [60, 61, 62, 63].

### 2.3.4. Summary of Evidence

Our review shows that existing evidence on digital discrimination in the sharing economy market is mainly geared towards documenting and addressing issues of racism and sexism as well as suggesting mitigation strategies for these issues (RQ₁). The majority
of these studies are published after 2015. Less evidence is available on other types of discrimination, such as classism or ableism, which are often investigated from a regulatory point of view [60, 61]. Our review also shows that most studies analyze discrimination in the domains of ridesharing, lodging, and freelancing (RQ₂). This calls for more research to investigate discrimination in other domains of the sharing economy such as food delivery or asset sharing [62, 85]. In terms of features and goals, our analysis revealed that discrimination concerns are commonly enabled by information available on user profiles and often find their way to reputation systems (RQ₃). A list of mitigation strategies are proposed to control for discriminatory behavior that might manifest through these features (RQ₄). These strategies range from preventive (e.g., withholding information, structured reviews, and trust cues) to corrective (e.g., bias correction and hiding older reviews) and even reactive (e.g., penalty for bias-based cancellations). A summary of these measures is listed in Table 2.5. Finally, we observe that inclusion, trust, safety, and accessibility are the main user goals often impacted by discrimination concerns.

### 2.4. Discussion and Impact

The research on digital discrimination has gained a significant momentum over the past four years. This can be attributed to the unprecedented widespread of the sharing economy systems and the general shift in society towards more equality and prosperity. As more research is conducted, it becomes harder for software engineers to keep up with this growing body of research. To address this limitation, our review in this chapter is intended to systematically synthesize existing evidence on digital discrimination from a

<table>
<thead>
<tr>
<th>User goal</th>
<th>Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trust</td>
<td>[2, 84, 81, 70, 82, 39, 49, 71, 72, 74, 73, 59, 77]</td>
</tr>
<tr>
<td>Safety</td>
<td>[22, 44, 83]</td>
</tr>
<tr>
<td>Accessibility</td>
<td>[60, 61, 62, 63]</td>
</tr>
<tr>
<td>Inclusion</td>
<td>all primary studies</td>
</tr>
</tbody>
</table>
Table 2.5. A summary of suggested design changes in the literature to mitigate discrimination.

<table>
<thead>
<tr>
<th>Design change</th>
<th>Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Withholding information</td>
<td>[7, 19, 70, 6, 46, 64]</td>
</tr>
<tr>
<td>Self-disclosure information</td>
<td>[44, 72, 19, 21]</td>
</tr>
<tr>
<td>Asset-based profile pictures</td>
<td>[70, 41]</td>
</tr>
<tr>
<td>Fully-automated matching</td>
<td>[51, 84]</td>
</tr>
<tr>
<td>Cashless payments</td>
<td>[48]</td>
</tr>
<tr>
<td>Mutual reviews</td>
<td>[70, 36]</td>
</tr>
<tr>
<td>Structured reputation systems</td>
<td>[76]</td>
</tr>
<tr>
<td>Explicit trust cue</td>
<td>[37, 54]</td>
</tr>
<tr>
<td>Hiding older reviews</td>
<td>[8, 77]</td>
</tr>
<tr>
<td>Bias correction</td>
<td>[55, 8, 86]</td>
</tr>
<tr>
<td>Bias-free rating elicitation</td>
<td>[55]</td>
</tr>
</tbody>
</table>

software design point of view. In general, our review shows that digital discrimination is far from being a simple problem. Such complex socio-technical phenomenon emerges from equally complex interactions between system features and their operational environment. Therefore, it is safe to say that there is no silver bullet for solving discrimination in the sharing economy market. However, *satisficing* solutions could be developed to mitigate the problem. These solutions can be inferred from existing interdisciplinary evidence which detects and documents discriminatory behavior in the sharing economy platforms through large scale field studies and controlled experiments.

The main objective of our SLR is to help software engineers to comprehend, communicate, and eventually integrate existing evidence on digital discrimination into their working systems. For instance, through our SLR, system designers can get insights into the complex interaction of features that could trigger or mitigate discrimination in their operational environment. Such information can be particularly important in agile environments where there is typically no time to research complex domain phenomena between product cycles.
In the long run, the impact of such work will extend to the entire population of the sharing economy users, targeting the deep racial and regional disparities in one of the fastest growing software ecosystems in the world.

The work presented in this chapter builds upon our previous work in this domain [10]. In our previous work, we analyzed a dataset of 667,806 tweets collected from the twitter feeds of six different sharing economy platforms. Our results showed that various forms of bias frequently appear in user feedback. We further conducted an ad-hoc literature review of 17 primary studies on digital discrimination. The results from our user feedback analysis as well as brief review were integrated into a partial model to capture the problem from a requirements engineering perspective. Our work in this chapter builds up on that perspective by conducting a more systematic literature review of existing literature on the problem.

2.5. Limitations and Threats to Validity

The study conducted in this chapter takes the form of a systematic literature review (SLR) [30, 87]. This method is commonly used for advancing the state-of-the-art in research and practice based on rigorous research, especially when the problem being investigated is inherently interdisciplinary. However, like most review-based studies, subjectivity threats can be raised about the quality of the quantitative and qualitative analysis performed by the reviewers as well as threats of missing related work. To mitigate these threats, we applied a set of well-known protocols for conducting evidence-based reviews [30]. Specifically, we searched multiple digital libraries for primary studies using a structured query and snowballing. Related studies were then identified using exclusion and inclusion criteria and synthesized using a systematic coding of evidence. To control for the validity of extracted evidence, the majority of primary studies considered in our review included some sort of a large scale field study or a controlled experiment that was conducted using a large number of observations. In addition, we used a grounded theory approach of open coding and
memoing to categorize extracted information. We believe that these actions helped to mitigate several potential threats affecting our study.

2.6. Conclusion

In this chapter, we proposed a new framework for modeling discrimination concerns in the sharing economy. In the first phase of our analysis, we systematically synthesized evidence from 58 interdisciplinary primary studies to extract information on the different types of discrimination concerns impacting the sharing economy platforms along with their mitigation strategies. The results showed that existing evidence is often related to issues of racism and sexism affecting the domains of ridesharing, lodging, and freelancing. The results also showed that discrimination concerns are commonly associated with the features of user profiles and reputation systems. These concerns are partially mitigated by a variety of design strategies that are introduced to prevent offline forms of systematic bias from transitioning online. Our review also showed that inclusion, trust, safety, and accessibility are the main user goals commonly intertwined with concerns of digital discrimination.

Our work in this chapter is intended to facilitate tasks of evidence-based software engineering in the sharing economy app development. In the future, we will seek to advance this line of work by integrating our synthesized evidence into actual working the sharing economy prototypes. This will enable us to investigate the impact of implementing some of the identified mitigation strategies in practical settings and objectively measure their success, or failure, in countering digital discrimination.
Chapter 3. User Perspective

In this chapter, we quantitatively and qualitatively analyze a large dataset of user feedback, collected from the Twitter feeds of eight popular sharing economy platforms. Our objective is to examine the distribution and types of digital discrimination in the online feedback of users of sharing economy platforms.

3.1. Introduction

Existing research on digital discrimination relies on direct user surveys (questionnaires and interviews) and field studies to identify discrimination concerns among sharing economy users [7, 8, 12]. However, these data collection methods are extremely costly, and the sample size, or response rate, are often limited by factors such as the geographical area covered, number of subjects surveyed, and number of platforms studied. To overcome these limitations, in our research, we propose to exploit the social media platform Twitter, as a source of online software user feedback. Previous work showed that Twitter has become a very active channel of communication between software developers and their end-users [88]. This can be particularly observed when the problem is of a social nature. For example, a search for discriminate AND (Uber OR Airbnb) returns tweets sighting incidents of discrimination over these platforms. In fact, the hashtag #AirbnbWhileBlack has become the main place for reporting and highlighting potential racial bias on the rental app Airbnb.

3.2. Data Scraping and Analysis

To conduct our analysis, we collected tweets related to eight main players in the sharing economy market. These systems cover the domains of ridesharing (Uber and Lyft), lodging (Airbnb and Couchsurfing), food delivery (Doordash and UberEats), and freelancing (TaskRabbit and Fiverr). Our data collection process extended over the period of two full months, from November 1st to December 31st, 2019. The data was collected using the Twitter Search API, considering only English tweets that contained the names of any of the eight systems included in our analysis. In total, 667,806 tweets were collected.
The main task after collecting our dataset is to locate our specific tweets of interest (discrimination related tweets). Several automated solutions have been proposed in the RE literature for mining Twitter data [88]. However, the majority of these solutions are proposed to classify tweets into generic maintenance tasks, such as feature requests and bug reports, with limited support for detecting issues of special nature, such as discrimination [89, 90]. In fact, finding such specific issues in large amounts of Twitter data has been described as finding a needle in a haystack [89].

To overcome this problem, in our analysis, we follow a snowballing approach. Snowballing is a commonly used strategy for exploratory data collection. This strategy starts with identifying an initial set of core strings (seeds) that are used for the first search query. Once the initial set of artifacts is located and examined, the search query is updated with new relevant terms acquired from the set, and another round of search is performed. The process continues until no new or relevant artifacts are found. Snowballing is commonly used in research as a reliable method for achieving high recall rates in tasks such as systematic literature reviews (SLRs) [69], search keyword identification [91], and Twitter data analysis [92].

To collect discrimination-related tweets, we began our search with the seed discr, which is the stem of the word discrimination. Stemming is necessary to count for the different morphological variants of the word (e.g. discriminate, discriminating, discriminated, and discrimination). In addition, the stems bigot for bigotry and prejud for prejudice were included since they often appear in English dictionaries as synonyms for the word discrimination [93]. Based on these seeds, we located our initial set of tweets.

Three researchers examined these tweets independently, following a systematic coding process to classify them into discriminatory and non-discriminatory tweets and to extract keywords for the subsequent search query [94]. Specifically, an initial meeting was held to discuss the common types of discrimination in today’s society. Then, for each tweet, each researcher had to answer three main questions a) does the tweet describe a discrimination
incident? \textit{b)} what is the broad type of the discrimination concern raised in the tweet? and \textit{c)} are there any other keywords that are strongly associated with the identified concern?. The coding process was carried over multiple sessions to avoid any fatigue issues and to ensure the integrity of the data [69]. A meeting was then held after the end of the coding phase to compile researchers' answers and to resolve coding issues. Such issues included conflicts in the classification of some types of concerns and missing discrimination-indicative words.

The set of identified keywords were picked based on how likely they would indicate discrimination. For example, in the tweet “\textit{... my argument is UberPool should be accessible for all customers}” the keyword \textit{discrimination} was used with the word \textit{accessible}. Since the user is complaining about the lack of accessibility, the stem \textit{accessib} was included in the set of indicator keywords for the next round of search. Extracted indicator words were then used to expand the query and the process was repeated for three rounds, until no more new keywords/tweets were found. At the end of this process, 22 unique discriminatory words or phrases (e.g. \textit{service dog}) were extracted from the dataset. Table 3.1 displays the extracted words for the three rounds of snowballing and the number of tweets obtained for each word. The included tweets are in parentheses.

It is important to point out that a large percentage of returned tweets were excluded from the final dataset for a variety of reasons. For instance, we did not include any tweets that were not tied to a user’s general or specific experience. For example, the tweet “\textit{Airbnb Works To Clean Up Its Reputation For Racial Discrimination In New 3-Year Report https://t.co/MRtHV07jjv}” was not included because the tweet was mainly publicizing a news article about discrimination over Airbnb. Another type of excluded tweets included tweets that were unrelated to discrimination to begin with. For example, the tweet “\textit{@wjxt4 No, that’s the parents discretion}” was returned as a possible match due to the presence of the stem \textit{discr} from the word \textit{discretion}, not \textit{discrimination}. 29
Table 3.1. The results of the snowballing tweet analysis.

<table>
<thead>
<tr>
<th>Round</th>
<th>Terms</th>
<th>Uber</th>
<th>Lyft</th>
<th>Airbnb</th>
<th>CouchSurfing</th>
<th>TaskRabbit</th>
<th>Fiverr</th>
<th>DoorDash</th>
<th>UberEats</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st</td>
<td>descr (discrimination), prejud (prejudice), bigot (bigotry)</td>
<td>280(22)</td>
<td>16(10)</td>
<td>104(21)</td>
<td>2(0)</td>
<td>0(0)</td>
<td>0(0)</td>
<td>8(6)</td>
<td>18(2)</td>
</tr>
<tr>
<td>2nd</td>
<td>car seat, raci (racism, racist), race, deaf, access, gender, disab (disability, disable), service animal, dog, gay, sex (sexism, sexist), lgbt, elder, wheelchair, religi (religious, religion)</td>
<td>1816(127)</td>
<td>228(48)</td>
<td>470(59)</td>
<td>4(0)</td>
<td>9(1)</td>
<td>5(0)</td>
<td>157(11)</td>
<td>184(1)</td>
</tr>
<tr>
<td>3rd</td>
<td>minor, handicap, trans, infant</td>
<td>126(0)</td>
<td>17(1)</td>
<td>8(4)</td>
<td>1(0)</td>
<td>1(0)</td>
<td>0(0)</td>
<td>7(0)</td>
<td>7(1)</td>
</tr>
</tbody>
</table>
3.3. Results and Conclusion

In general, four types of discrimination were detected in our data: racism, sexism, ableism, and parental. Racism concerns were detected in tweets such as "@Airbnb you have racist host users who deny stay to guests that are not clearly white.". Ableism concerns were detected in tweets such as "@DoorDash please tell your drivers to consider disabled customers before telling them come out to retrieve their orders. It’s highly offensive.". Sexism issues were detected in tweets such as "@gem_zam @doxiebaby I literally was refused an Uber 2 weeks ago b/c the driver didn’t want a gay passenger. Many lgbtq can not hide behind anything some can, and there is privilege in that, but many if not most of us cannot.". Discrimination against parents were detected in tweets such as "@Uber_Support it’s really important my account is opened. It’s the only way I have to travel in Canada. Especially after I just gave birth! Is there a discrimination over woman having infants? Of course I have a car seat! But that sounds not the issue!!". In addition, there were several tweets that reported discrimination incidents without specifying exactly what type of discrimination took place, such as “@Airbnb @AirbnbHelp Why close my complaint on discriminatory behavior by host without a proper resolution? After accepting payment, host cancels the booking on discriminatory grounds. Is this what one has to expect from #AirBnB?.”

In terms of platforms, we observed that the ridesharing services, Uber and Lyft, suffered from the most cases of discrimination, followed by the lodging service Airbnb. In fact, these results were expected given the popularity of these services over other services such as Fiverr or Couchsurfing. We also observed that food delivery platforms had instances of discrimination, however, such tweets were not as common as in ridesharing data. These observations suggest that user data for these platforms should be collected over longer periods of time in order to increase the chances of capturing discrimination-related tweets. A breakdown of discrimination-related tweets (number of tweets) per platform is provided in Table 3.2.
Table 3.2. Number of discrimination tweets per each sharing economy platform.

<table>
<thead>
<tr>
<th>Type</th>
<th>Uber</th>
<th>Lyft</th>
<th>Airbnb</th>
<th>Couchsurfing</th>
<th>TaskRabbit</th>
<th>Fiverr</th>
<th>DoorDash</th>
<th>UberEats</th>
</tr>
</thead>
<tbody>
<tr>
<td>Racism</td>
<td>84</td>
<td>33</td>
<td>31</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Ableism (Disability)</td>
<td>29</td>
<td>7</td>
<td>33</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Sexism</td>
<td>24</td>
<td>15</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Parental</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>9</td>
<td>3</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>1</td>
</tr>
</tbody>
</table>

In summary, the results of this analysis show that discrimination concerns do exist and they often get reported over social media. However, these concerns tend to be scarce and buried within large amounts of irrelevant tweets as well as vary in their quantity and intensity among different platforms. For instance, given the commercial nature of sharing economy platforms, their Twitter feeds tend to be overloaded with spam. Furthermore, some of the popular sharing economy services have become household names, or even used as verbs (e.g., “I’m going to Uber to work today”). Therefore, isolating tweets that actually raise discrimination concerns among these tweets that simply mention the name of the service can be a very laborious task. We further noticed that a large number of tweets were very brief in describing incidents of discrimination with no details about the incident (e.g., “My Uber driver mad sexist Jesus Christ”). This can be attributed to the nature of Twitter as a micro-blogging service that does not allow messages longer than 280 characters.

Finally, we observed that the majority of discrimination-related tweets reported the experience of consumer (e.g., renters or riders), with only a small percentage reporting issues from the service provider side (e.g., hosts or drivers). This emphasizes the need for using other types of data collection methods (e.g., surveys and field studies) in order to capture the concerns of all types of users.

3.4. Threats to Validity

The study conducted in this paper suffers from several methodological constraints that might jeopardize the validity of our findings. A main threat to the validity of our study stems from the fact that our data was collected from Twitter and only for a relatively limited
period of time. A recent report by the Pew Research Center has shown that most Twitter users rarely tweet, and the most prolific 10% create 80% of traffic among adult users. In the U.S., only 22% of American adults use Twitter, and this segment tends to be younger, more highly educated and wealthier than the general public. The report also states that Twitter users are more likely to be sensitive to issues of racial discrimination [95]. Twitter might also conceal sampling bias given that the demographic of users (e.g., gender, age, and location) is unknown. However, as mentioned earlier, our goal in this Chapter is to develop a preliminary perspective of the problem. Twitter, as a social media platforms, is expected to provide a low-cost preliminary evidence given that discrimination is inherently a social problem. Nonetheless, we acknowledge the fact that data collected over longer periods of time and from other channels of feedback, such as app store reviews, online blogs, and direct user surveys, are necessary to achieve a better coverage of the problem and eliminate sampling bias, especially for smaller platforms that do not typically receive a large number of tweets.
Chapter 4. Policies

Recent research has exposed a serious discrimination problem affecting applications of the sharing economy, such as Uber, Airbnb, and TaskRabbit. To control for this problem, several sharing economy apps have crafted a new form of usage policies, known as non-discrimination policies (NDPs). These policies are intended to outline end-users’ rights of equal treatment and describe how acts of bias and discrimination over sharing economy apps are identified and prevented. However, there is still a major knowledge gap in how such non-code artifacts can be formulated, structured, and evolved. To bridge this gap, in this chapter, we introduce a first-of-its-kind framework for analyzing and evaluating the content of NDPs in the sharing economy market. Our analysis is conducted using a dataset of 108 sharing economy apps, sampled from a broad range of application domains. Our results show that, a) most sharing economy apps do not provide a separate NDP, b) the majority of existing policies are either extremely brief or combined as sub-statements of other usage policies, and c) most apps do not provide a clear statement of how their NDPs are enforced. Our analysis in this chapter is intended to assist sharing economy app developers with drafting and evolving more comprehensive NDPs as well as help end-users of these apps to make more informed socioeconomic decisions in one of the fastest growing software ecosystems in the world.

4.1. Introduction

In response to discrimination concerns, the sharing economy developers have started rolling out a new form of policies for addressing potential issues of discrimination affecting their apps. A policy, in general, serves as a legally binding contract between apps and their end-users [96]. For instance, popular app marketplaces demand apps to provide a privacy policy to specify the types of information they collect about their users and outline how such information is being used, protected, and shared [96]. Similarly, non-discrimination policies (NDPs) are expected to determine the app’s stance on discrimination and outline how acts of discrimination over the app are identified and handled. Privacy policies have
received significant attention in the Software Engineering literature [97, 98]. This line of research aims to assess the quality of privacy policies as well as gauge best practices for drafting them. However, there is a widespread lack of knowledge about how NDPs can be structured. This can be attributed to the fact that NDPs are non-code artifacts. Creating and evolving such artifacts about a complex socio-technical phenomenon such as digital discrimination often fall outside of developers’ expertise.

To address this knowledge gap, in this chapter, we develop a framework for systematically analyzing and evaluating the content of NDPs in the sharing economy market. Our analysis is conducted using a dataset of 108 sharing economy apps, sampled from a broad range of application domains. The objectives of the proposed framework are to a) assist sharing economy app developers with drafting and evolving more comprehensive and less ambiguous NDPs, and b) help end-users of sharing economy apps to make more informed socioeconomic decisions in the sharing economy market, either as service providers or receivers.

The remainder of this chapter is organized as follows. Section 4.2 motivates our work and discusses our research questions. Section 4.3 describes our data collection process. Section 4.4 describes our NDP quality assessment framework. Section 4.5 presents our results. Section 4.6 discusses our key findings and their impact as well as the main limitations of our study. Finally, Section 4.7 concludes the chapter and discusses our future work.

4.2. Motivation and Research Questions

In this section, we review existing evidence on the problem of digital discrimination, motivate our work, and discuss our research questions.

4.2.1. Digital Discrimination

The problem of digital discrimination in online sharing economy markets has been well-documented in recent years. Numerous large-scale surveys and field studies have provided significant evidence on various forms of systematic bias across almost all application domains of sharing economy, including discrimination based on ethnic background (racism),
gender or sexual orientation (sexism), and physical appearance (ableism) [6, 11, 7, 12, 13, 14]. For instance, Ge et al. [6] hired research assistants of different racial backgrounds to request UberX rides. The authors found that the waiting times for Black riders were significantly longer. In addition, more cancellations were observed against Black riders than their White counterparts. In another study, Moody et al. [11] surveyed 1,100 of UberPOOL and Lyft riders. The results showed that White passengers that lived in predominantly White communities were more likely to discriminate against passengers of other races.

Edelman et al. [7] examined racial discrimination over the lodging platform Airbnb. The authors reported that applications from guests with distinctively Black names were 16% less likely to be accepted relative to identical guests with distinctively White names. Discrimination in the lodging business has also been observed against members of the LGBT community. For example, Ahuja and Lyons [12] analyzed Airbnb hosts’ responses to LGBT accounts. The results showed that hosts were more likely to not reply at all rather than replying “no” to male-male pairs inquiring about room availability. Ableism (discrimination against people with disabilities) was also reported over Airbnb. For instance, in a randomized field experiment of 3,847 lodging requests, Ameri et al. [63] found that hosts were less likely to approve requests from travelers with blindness, cerebral palsy, dwarfism, or spinal cord injury than to approve travelers without disabilities.

Patterns of digital discrimination have also been observed in the freelancing domain. Thebault et al. [13] surveyed workers on TaskRabbit from the Chicago metropolitan area. The authors found that requests from customers in the socioeconomically disadvantaged South Side area were less likely to be accepted. Hannák et al. [8] analyzed worker profiles on TaskRabbit and Upwork. The results showed that there was a significant bias against White women and Black men on both platforms. In another study, Foong et al. [56] collected self-determined hourly bill rates from the public profiles of 48,019 workers in the U.S. (48.8% women) on Upwork. The authors found that the median woman on Upwork requested only 74% of what the median man requested in hourly bill rate. Another study by
Barzilay and Ben-David [14] showed that women’s average hourly rates on P2P freelancing platforms were about two-thirds of men’s rates. Such gaps persisted even after controlling for experience, educational background, and hours of work.

4.2.2. Motivation and Research Questions

Policies have long been used as legally-binding usage contracts between software platforms and their end-users [96]. For instance, privacy policies are used by app developers to communicate their data collection and sharing practices with their end-users as well as to comply with privacy legislation around the world. These policies have generated significant research interests in recent years [99]. Privacy policy research is primarily focused on detecting violations of the claims made in the policy [100, 101], evaluating the readability and comprehensibility of policies [100, 101], and mining their content for software privacy requirements [102, 97, 98]. NDPs, on the other hand, have received considerably less attention in both research and practice. This can be attributed to the fact that digital discrimination is an inherently complex phenomenon that is often enabled by equally complex interactions between sharing economy apps’ features, their end-users, and operational environments. Therefore, drafting NDPs that are tailored to address the specific types of bias affecting different application domains can be a very challenging and time-consuming process.

To address these limitations, in this chapter, we conduct a first-of-its-kind study to analyze NDPs in the Digital Sharing Economy. Our work aims to a) study the prevalence of NDPs in the sharing economy market, b) propose a framework for systematically analyzing the content of these policies, and c) use that framework to assess the quality of existing NDPs. Our work is intended to spread awareness of digital discrimination and provide app developers, either maintaining sharing economy apps or developing new ones, with systematic guidelines to draft high quality NDPs and evolve such non-code artifacts with minimum overhead. Moreover, providing complete and structured NDPs can help sharing economy app users to make more optimized socioeconomic decisions when it comes to
navigating the landscape of existing sharing economy platforms. To guide our analysis, we formulate the following research questions:

- **RQ1:** *How prevalent are NDPs in the sharing economy market?* Under this research question, we investigate the prevalence of anti-discrimination policies among sharing economy apps. This type of analysis aims to explore the state-of-practice in the different application domains of sharing economy when it comes to NDPs.

- **RQ2:** *Can the quality of existing NDPs be systematically evaluated?* Under this research question, we seek to develop a systematic framework for analyzing the content of existing NDPs as well as assess their quality.

- **RQ3:** *How detailed and informative are existing NDPs?* Under this research question, we examine the quality of information provided in existing NDPs. Our objective is to determine a set of quality standards that can be used by new sharing economy apps, or existing apps with no policies, to draft and evolve their own NDPs.

4.3. Data Collection

In this section, we describe our data collection process, including selecting apps to be included in our dataset, categorizing these apps, and collecting their NDPs.

4.3.1. Dataset

Recent statistics estimate that there are thousands of active sharing economy platforms listed on popular mobile app marketplaces. However, only a handful of these apps are typically investigated in digital discrimination research. Such apps include Uber and Lyft from the domain of ride-sharing, Airbnb from the lodging domain, and Upwork and TaskRabbit from the domain of freelancing [6, 11, 7, 12, 13, 14]. These apps operate in large geographical areas and have massive user bases, thus, discrimination concerns are more likely to manifest over them rather than smaller ones. Based on these observations,
for a sharing economy platform to be included in our analysis, it has to meet the following criteria:

1. A platform must facilitate some sort of a P2P connection and include the sharing of some sort of a resource, such as an asset (e.g., an apartment, car, electric drill, etc.) or a skill (e.g., plumbing, hair styling, coding, etc.).

2. A platform must have an app on Google Play or the Apple App Store. App stores provide various metrics that can help us to locate popular apps, such as the number of app reviews, stars, and their download statistics.

3. A platform must be located and/or have a substantial presence in the US. The U.S. Civil Rights Act of 1964 prohibits discrimination based on race, sex, religion, nationality, or sexual orientation. By focusing on the US market, we ensure that our selected apps operate in a country where discrimination is prohibited by law.

With these criteria in place, we searched for apps to be included in our dataset. Our data collection took place between January and February of 2021. We started by seeding our dataset with Uber, Lyft, Airbnb, Upwork, TaskRabbit, and Fiverr. Existing literature has provided a significant evidence of discriminatory behavior affecting these apps. We then conducted a Google search using the query: (sharing OR shared OR gig) AND economy AND (platforms OR apps OR systems). We examined the first 10 pages of the search results and added 72 new platforms that matched our inclusion criteria. We then used the similar feature on Google Play and the Apple App Store to locate any apps we missed through the Google search. Specifically, we examined the list of similar apps resulting from searching app stores for each of our 72 apps. Lightweight snowballing was then used to add any major apps that we might have missed. Apps were iteratively added until no more new apps that satisfied our inclusion criteria were located. In total, 108 unique apps were included in our dataset. Descriptive statistics of our dataset are provided in Table 4.1.
Table 4.1. Descriptive statistics for the 108 apps in our dataset.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Mean</th>
<th>Median</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>App Store Rating</td>
<td>4.23</td>
<td>4.60</td>
<td>1.60</td>
<td>4.90</td>
</tr>
<tr>
<td>Google Play Rating</td>
<td>3.86</td>
<td>3.90</td>
<td>2.00</td>
<td>4.90</td>
</tr>
<tr>
<td>App Store # of Reviews</td>
<td>201K</td>
<td>2.4K</td>
<td>2</td>
<td>8.9M</td>
</tr>
<tr>
<td>Google Play # of Reviews</td>
<td>134K</td>
<td>1.3K</td>
<td>7</td>
<td>7.91M</td>
</tr>
<tr>
<td>Google Play # of Installs</td>
<td>6.9M</td>
<td>100K</td>
<td>1K</td>
<td>500M</td>
</tr>
</tbody>
</table>

4.3.2. App categorization

The Apple App Store and Google Play classify apps into generic categories of loosely related functionalities. These categories are often ambiguous (too generic) or straight-up misleading [103, 104]. For example, both Uber and Airbnb are categorized under the Travel category in the Apple App Store and DoorDash is classified under the Food&Drink category. This type of generic categorization does not provide enough information about the specific application domains of apps. To overcome this limitation, we begin our analysis by re-classifying apps in our dataset into more fine-grained categories of sharing economy application domains.

While automated app classification techniques are available [103, 104], given the relatively small size of our dataset, we conducted the classification manually. In particular, three judges, all with graduate degrees in Software Engineering and an average of three years of industrial experience, independently examined the description of each of our apps available on the Apple App Store and Google Play as well as each app’s official web-page. Categories of apps were recorded as they emerged in the text. We used memoing to keep track of the reasoning behind each suggested category. Axial coding was then used to consolidate individual categories into more abstract categories [105]. For example, the categories of food delivery and grocery delivery were merged into a single Delivery category and boat-sharing and bike-sharing were merged into asset-sharing. Generated categories were then iteratively revised until no more categories were found. By the end of our clas-
Figure 4.1. The application domains of the sharing economy apps in our dataset.

In our classification process, six main categories of sharing economy apps, shown in Fig. 4.1, have emerged. These categories can be described as follows:

- **Skill-based.** These apps facilitate the sharing of personal skills (hiring labor). Specific examples include the baby sitting apps Sittercity and Urbansitter, the tutoring apps Verbling, Codementor, and Classgap, and the freelancing apps Fiverr and Upwork.

- **Delivery.** Under this category, we include apps which enable users to utilize their vehicles to deliver goods to other users. Examples of apps in this category include UberEats, Grubhub, and Shipt for grocery and food delivery, and DriveMatch, uShip, and Dolly for hiring delivery drivers.

- **Ride-sharing.** This category includes apps which allow their users to share rides, such as carpooling and driver/rider connections. Examples of apps in this category include traditional ride-sharing services, such as Uber, Lyft, and Via, as well as more specialized platforms, such as HopSkipDriver for children transportation, Veyo for medical transportation, and Wingz for hiring a driver.

- **Asset-sharing.** Under this category, we include any app which enables users to share their assets. Specifically, the resource being shared is a physical resource (e.g., a vehicle or an electric drill), not a person’s time or skills (e.g., a driver or electrician).
Examples of apps under this category include the car sharing apps Turo and HyreCar, the boat sharing apps Get-MyBoat and Boatsetter, the bike sharing app Spinlister, and the RV sharing apps RVezy and Outdoorsy.

- **Lodging.** This category contains renting and short-term accommodation services such as Airbnb, Vrbo, and Misterbnb as well as space-sharing for storage (Neighbor), events (Splacer), and even parking (ParqEx).

- **Other.** Although our objective was to classify all apps into the main general categories, two apps in our dataset were too niche-oriented to warrant a creation of a separate category. These apps are Prosper for lending and borrowing money and Kickstarter, a platform for crowdfunding various projects.

### 4.3.3. Policy collection

To answer our first research question, we collect the NDPs of the apps in our dataset. Unlike privacy policies, mobile app marketplaces do not enforce NDPs, therefore, locating such policies can be a challenging task. For instance, most privacy policies are often titled *Privacy Policy*, however, NDPs are titled differently, including titles such as, *non-discrimination*, *anti-discrimination*, or *inclusion statement*. To locate such policies, we explore the website of each app as well as the app itself. Any web pages or app screens that address discrimination are collected as a potential NDP.

To identify these pages, we utilized Google’s search operators to search apps’ websites directly using the query `site: <app website> AND (discrimination OR <discrimination types (Table 4.2)>)).` Table 4.2 lists the main acts of discrimination as described by the U.S. Equal Employment Opportunity Commission. These acts commonly appear in diversity and social justice literature [32]. For any app that we could not locate a policy, we performed a manual search of its website. Our search exposed three categories of apps when it comes to NDPs. These categories include:
Table 4.2. Most common types of discrimination in the literature.

<table>
<thead>
<tr>
<th>Type</th>
<th>Discrimination against:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Racism</td>
<td>Ethnicity, color, or nationality.</td>
</tr>
<tr>
<td>Sexism</td>
<td>Gender or sexual orientation.</td>
</tr>
<tr>
<td>Ableism</td>
<td>Physical, sensory, or intellectual disability.</td>
</tr>
<tr>
<td>Parental</td>
<td>Parents with children or pregnant women.</td>
</tr>
<tr>
<td>Ageism</td>
<td>Older or younger people.</td>
</tr>
<tr>
<td>Religious</td>
<td>Perceived religion or a set of beliefs.</td>
</tr>
<tr>
<td>Classism</td>
<td>Particular social class.</td>
</tr>
</tbody>
</table>

- **Separate policy.** This category includes apps which maintain a separate NDP that is provided on its own separate page. In total, 16 apps had a separate NDP.

- **Combined policy.** Nine apps in our dataset combined their NDP with other usage policies, such as sexual harassment policies, community guidelines, code of conduct, or even the Terms of Service (ToS) of the app.

- **No policy.** For the majority of apps (79) in our dataset, we were either unable to locate a policy, or only located a generic one-line anti-discrimination statement that was provided in the ToS of the app. Some apps provide some sort of a statement on diversity or commitment to diversity. These statements typically take the form of a blog post rather than being a policy with rules and implications. For example, Gopuff, a delivery app, published a commitment to creating more equal and just future in response to the death of George Floyd.

The distribution of these three categories of NDPs over our categories of sharing economy application domains is shown in Fig. 4.2. In general, to answer *RQ*$_1$, we can safely say that the majority of apps in our dataset do not provide NDPs. We found that some apps merge their NDPs with other policies, while only a few of the apps publish a separate NDP.
4.4. Quality Assessment of NDPs

In this section, we propose a framework for assessing the quality of NDPs in the sharing economy market. The process of policy assessment is typically conducted manually, following a systematic process that checks the content of the policy against a set of predefined quality measures [106, 107, 108, 109, 110]. These measures range from simple quantitative metrics, such as the length of the policy [111], to more complex measures, such as its readability and compliance with regulations [112]. To generate such a protocol, we rely on two sources of information:

- **Nondiscrimination regulations.** The U.S. Equal Employment Opportunity Commission (EEOC) suggests an outline of topics that US-based employers should include in their NDPs\(^1\). While these guidelines focus on discrimination against employees (rather than end-users), they can be used to establish the structure of NDPs. For example, the EEOC guidelines state that NDPs should include specific types of discrimination, a reporting mechanism, and consequences of violating the policy.

- **Privacy policy assessment protocols.** Existing protocols for evaluating privacy policies can serve as a baseline, or a reference, to develop an evaluation protocol for NDPs. Such protocols include a set of measures that can be directly inferred from

---

\(^1\) [https://www.eeoc.gov/employers/small-business/general-non-discrimination-policy-tips](https://www.eeoc.gov/employers/small-business/general-non-discrimination-policy-tips)
the policy. Typically, evaluators are provided with a set of questions to help them evaluate the content of the policy based on the predefined measures [112, 113, 106].

Based on these two sources of information, we design a protocol for evaluating NDPs in the sharing economy market. The specific measures of our protocol, along with their descriptions and their associated evaluation questions are provided in Table 4.3. In general, our measures can be divided into a set of automatically calculated measures, including the policy’s length and readability, and manually determined measures, including the types of discrimination mentioned in the NDP, the number of examples provided, references to legislation, and whether the policy mentions enforcement and ramifications mechanisms.
Table 4.3. Assessment measures of NDP quality.

<table>
<thead>
<tr>
<th>No</th>
<th>Measure</th>
<th>Description</th>
<th>Questions for evaluators</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Name (N)</td>
<td>The name of the policy is the title of the document the policy is listed under. Separate policies with well-defined titles are more easily accessible, thus can be considered higher in quality [111].</td>
<td>Determined during policy collection.</td>
</tr>
<tr>
<td>2</td>
<td>Length (L)</td>
<td>The length of the policy (number of words) can be used as a basic measure of its quality. Intuitively, longer policies are assumed to be more detailed [111]</td>
<td>Measured automatically as the number of words.</td>
</tr>
<tr>
<td>3</td>
<td>Readability (FRE.)</td>
<td>Readability is another measure that is commonly used to assess the quality of policies [111, 114, 115]. The more readable the policy, the more accessible it is for the casual user.</td>
<td>Calculated automatically using the Flesch Reading Ease (FRE) metric. [116]</td>
</tr>
<tr>
<td>4</td>
<td>Types (T)</td>
<td>Discrimination in sharing economy apps can take many forms (Table 4.2). Therefore, a policy that explicitly mentions more of these types is considered higher in quality, or more comprehensive. The US EEOC states that discrimination based on race, color, religion, sex (including pregnancy, sexual orientation, or gender identity), national origin, disability, age or genetic information (including family medical history) is illegal.</td>
<td>How many specific types of discrimination does the policy mention?</td>
</tr>
</tbody>
</table>

(table cont’d.)
<table>
<thead>
<tr>
<th>No</th>
<th>Measure</th>
<th>Description</th>
<th>Questions for evaluators</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Examples (Ex.)</td>
<td>A policy which provides examples of specific types of discriminatory behavior that might affect the app is considered to be higher in quality. Examples are used to demonstrate what actions might be classified as discriminatory. In policy analysis, examples are considered an important instrument to communicate policy practices with the casual user [108].</td>
<td>Does the policy provide any examples of discriminatory behavior? How many examples are provided?</td>
</tr>
<tr>
<td>6</td>
<td>Legislation (Lg.)</td>
<td>This criterion assesses whether a policy contains references to existing anti-discrimination regulations in the judicial area in which the app operates. For example, Internet privacy policies are often assessed based on their compliance with existing privacy regulations [106], such as the Federal Trade Commission’s Fair Information Practices guidelines [108].</td>
<td>Does the policy refer to any existing legislation?</td>
</tr>
<tr>
<td>7</td>
<td>Enforcement (En.)</td>
<td>A policy which lists the measures (functional or non-functional) taken by the app to mitigate discrimination is considered higher in quality [108]. In fact, the US EEOC states that a NDP should explain how employees can report discrimination. These types of mechanisms also include methods for reporting incidents of policy violation.</td>
<td>Does the policy list any features or protocols that the app implements to mitigate discrimination? Is there a reporting mechanism in place?</td>
</tr>
<tr>
<td>8</td>
<td>Ramifications (Rmf.)</td>
<td>A policy which mentions the ramifications for discriminatory behavior is considered more comprehensive. The US EEOC states that a NDP should describe the consequences of violating the policy.</td>
<td>Does the policy mention the types of actions (penalties) to be imposed on policy violators?</td>
</tr>
</tbody>
</table>

(table cont’d.)
Once our review protocol was defined, we printed out the NDPs collected for our apps. Each of our three judges went through each NDP independently, answering the questions related to the criteria from (4-8) in Table 4.3. Results were then compiled and summarized in Table 4.4, with three largest values in each numerical column highlighted. Overall, given the specific nature of our questions, only a few coding errors (inaccuracies in answering some of the questions) were detected and corrected.

4.5. Results and Analysis

In this section, we discuss the results of applying our evaluation protocol in Table 4.3 to the NDPs in our dataset.

4.5.1. Policy Name

Our results show that NDPs are named differently by different apps. Titles, such as Anti-Discrimination Policy and Non-Discrimination Policy are common. However, we found more variations of these titles, such as Zero Tolerance Policy, Deactivation Policy, Inclusion Policy, and more. In general, these variations can impact the accessibility of NDPs negatively [111]. This became clear during our data collection as we had to resort to a sophisticated Google query to retrieve the NDPs of our apps (Sec. 4.3.3).

4.5.2. Length and Readability

The Flesch Reading Ease (FRE) [116] is a popular metric used to assess the readability of text. The value of FRE ranges from 0 to 100, where a higher score indicates that the text is easier to read. The metric is calculated by the following formula:

\[
206.835 - 84.6 \times \frac{\text{# of syllables}}{\text{# of words}} - 1.015 \times \frac{\text{# of words}}{\text{# of sentences}}
\]  

(4.1)

The core idea behind FRE is that longer words and longer sentences are more difficult to comprehend. Therefore, FRE penalizes texts with a high number of syllables per word and a high number of words per sentence. Fig. 4.3 shows an example of how FRE can be calculated for a single sentence with 23 syllables and 11 words.
Dis/crim/i/na/tion of a/ny kind is not tol/er/at/ed in the Tu/ro com/mu/ni/ty.

\[ FRE = 206.835 - 84.6 \times \frac{23}{11} - 1.015 \times \frac{11}{1} = 18.78 \]

Figure 4.3. An example of FRE calculation for a text with a single sentence.

Figure 4.4. Length and readability of the NDPs in our dataset.

FRE is commonly used in policy assessment research [111, 114, 115]. It is important to note that this metric is only suitable for longer texts. Therefore, we calculated FRE only for NDPs with 100 words or more. The distribution of length and readability scores over our NDPs with length \( \geq 100 \) are presented in Fig. 4.4. Our results show that the average FRE for the policies in our dataset is 23.74. This level indicates that the text is difficult to read, best understood by college graduates. The apps Misterb&b, Spareroom, and Sittercity have the highest readability scores, while Roadie and Thumbtack have the lowest scores. In terms of length, Airbnb, GoShare, and Turo have the longest, thus more detailed policies. Uber’s and Lyft’s NDPs were surprisingly short (134 and 97 words respectively).
4.5.3. Types

Our annotation shows that most policies list a large number of discrimination types in their NDPs. TaskRabbit, in particular, refers to 17 different types, including racism, color, ancestry, national origin, religion, creed, age, sex, gender, physical or mental disability, medical condition, genetic information, marital or civil partner status, military or veteran status. The policy even provides more sub-types of discrimination, such as, “gender (including pregnancy, childbirth, breastfeeding or related medical conditions).” On average, NDPs in our dataset mention 10 types of discrimination per policy. Racism, national origin, disability, religion, age, gender identity, and marital status are the most frequent (Fig. 4.5).

4.5.4. Examples

Our manual annotation shows that examples are not common in NDPs. Airbnb, Turo, and Neighbor provide the most comprehensive set of examples, described in the form of "user may not" scenarios that could take place while using the app. For instance, Airbnb policy states that, “Airbnb hosts may not decline a booking from a guest based on gender identity unless the host shares living spaces (for example, bathroom, kitchen, or common areas) with the guest”. Neighbor’s NDP provides some of the best examples in terms of quantity and quality. The app provides 15 examples of what is considered discriminatory behavior, such as, “Posts that assume someone is suspicious because of their race or ethnicity”. Turo is another app which provides thorough examples of discriminatory behavior, such as, “Turo hosts may not make assumptions about the guest’s ability to operate their vehicle.”

4.5.5. Legal

Only three apps in our dataset provide references to specific counter-discrimination legislation. GoShare’s NDP for example, states that, “A variety of federal, state, and local laws strictly prohibit such forms of discrimination, including Title VII of the Civil Rights Act of 1964, the Age Discrimination Act of 1967, and the Americans with Disabilities Act of 1990.” Some other apps provide a generic legal statement. For example, TaskRabbit’s
Figure 4.5. A frequency-based word cloud of the different types of discrimination mentioned in the NDPs of apps in our dataset.

NDP states that, “or any other basis protected by applicable laws in jurisdictions in which TaskRabbit operates (collectively referred to as a protected class).” Similarly, Upwork’s NDP states that, “we expect all clients and freelancers to comply at all times with the laws concerning discrimination and harassment.” Turo’s NDP refers to cases outside the United States and Canada, stating that, “hosts aren’t required to comply with the above policies if they violate local laws.” However, no references to any specific laws are made.

4.5.6. Enforcement and Ramifications

A total of 15 apps in our dataset describe a set of measures taken to enforce their NDPs, including the ramifications for violating the policy. In general, our analysis revealed two categories of enforcement mechanisms:

- **Monitoring.** Some apps indicate in their policies that they monitor the actions of their users to detect discriminatory behavior. For example, Airbnb and Neighbor state in their NDPs that they may suspend hosts who have demonstrated a pattern of rejecting guests from a protected class. Furthermore, listings over these apps are constantly checked for language contrary to their nondiscrimination policies. It is not clear, however, what constitutes a pattern of discrimination or what language is considered discriminatory.

- **Reporting.** Some apps use reporting mechanisms to enable their users to report any incidents of perceived discrimination or unlawful bias. For instance, Turo has a
"Support Form" for users to report issues of discrimination and GoShare provides a full procedure on how to report alleged cases of discrimination and harassment.

In terms of ramifications, most apps which provide an enforcement mechanism also provide a statement indicating that proven cases of frequent discrimination would result in removal from the app or suspending the user temporarily or indefinitely.
<table>
<thead>
<tr>
<th>Domain</th>
<th>NDP title</th>
<th>App</th>
<th>NDP type</th>
<th>L</th>
<th>FRE</th>
<th>T</th>
<th>Ex</th>
<th>Lg</th>
<th>En</th>
<th>Rnf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ride-sharing</td>
<td>Non-Discrimination Policy</td>
<td>Uber</td>
<td>Separate</td>
<td>134</td>
<td>18.78</td>
<td>9</td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>Anti-Discrimination Policies</td>
<td>Lyft</td>
<td>Separate</td>
<td>97</td>
<td>-</td>
<td>10</td>
<td>0</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>Anti-Discrimination Policy</td>
<td>Via</td>
<td>Separate</td>
<td>102</td>
<td>30.77</td>
<td>8</td>
<td>0</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>Zero Tolerance Policy</td>
<td>HopSkipDrive</td>
<td>Combined</td>
<td>37</td>
<td>-</td>
<td>11</td>
<td>0</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>Non-Discrimination Statement</td>
<td>Veyo</td>
<td>Separate</td>
<td>189</td>
<td>8.30</td>
<td>11</td>
<td>0</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td>Non-Discrimination Policy</td>
<td>Wingz</td>
<td>Separate</td>
<td>88</td>
<td>-</td>
<td>9</td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Asset-sharing</td>
<td>Non-Discrimination Policy</td>
<td>Turo</td>
<td>Separate</td>
<td>770</td>
<td>23.78</td>
<td>13</td>
<td>15</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Table 4.4. NDP content assessment results.
<table>
<thead>
<tr>
<th>Domain</th>
<th>App</th>
<th>NDP title</th>
<th>NDP type</th>
<th>L</th>
<th>FRE</th>
<th>T</th>
<th>Ex</th>
<th>Lg</th>
<th>En</th>
<th>Rmf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delivery</td>
<td>Doordash</td>
<td>Deactivation Policy</td>
<td>Combined</td>
<td>64</td>
<td>-</td>
<td>14</td>
<td>0</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Grubhub</td>
<td>Policy Against Sexual and Other Forms of Harassment</td>
<td>Combined</td>
<td>60</td>
<td>-</td>
<td>16</td>
<td>0</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>uShip</td>
<td>Code of Conduct</td>
<td>Combined</td>
<td>153</td>
<td>35.17</td>
<td>9</td>
<td>0</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>GoShare</td>
<td>Anti-Sexual Harassment and Anti-Discrimination Policy</td>
<td>Combined</td>
<td>1424</td>
<td>13.00</td>
<td>11</td>
<td>0</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Postmates</td>
<td>Anti-Discrimination Policy</td>
<td>Separate</td>
<td>72</td>
<td>-</td>
<td>11</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Roadie</td>
<td>Discrimination And Sexual Harassment Policy</td>
<td>Combined</td>
<td>480</td>
<td>12.10</td>
<td>13</td>
<td>0</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Instacart</td>
<td>Community Guidelines for Customers</td>
<td>Combined</td>
<td>35</td>
<td>-</td>
<td>11</td>
<td>0</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

(table cont’d.)
<table>
<thead>
<tr>
<th>Domain</th>
<th>App</th>
<th>NDP title</th>
<th>NDP type</th>
<th>L</th>
<th>FRE</th>
<th>T</th>
<th>Ex</th>
<th>Lg</th>
<th>En</th>
<th>Rmf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lodging</td>
<td>Airbnb</td>
<td>Nondiscrimination Policy</td>
<td>Separate</td>
<td>2012</td>
<td>34.70</td>
<td>8</td>
<td>22</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Misterb&amp;b</td>
<td>Anti-Discrimination Policy</td>
<td>Separate</td>
<td>420</td>
<td>52.40</td>
<td>5</td>
<td>5</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Vrbo</td>
<td>-</td>
<td>Separate</td>
<td>91</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Neighbor</td>
<td>Nondiscrimination Policy</td>
<td>Combined</td>
<td>572</td>
<td>30.90</td>
<td>8</td>
<td>15</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Spareroom</td>
<td>Fair Housing</td>
<td>Separate</td>
<td>399</td>
<td>44.33</td>
<td>14</td>
<td>5</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
</tr>
</tbody>
</table>

*(table cont’d.)*
<table>
<thead>
<tr>
<th>Domain</th>
<th>App</th>
<th>NDP title</th>
<th>NDP type</th>
<th>L</th>
<th>FRE</th>
<th>T</th>
<th>Ex</th>
<th>Lg</th>
<th>En</th>
<th>Rmf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Skill-based</td>
<td>Taskrabbit</td>
<td>Anti-Discrimination and Harassment Policy</td>
<td>Separate</td>
<td>403</td>
<td>0</td>
<td>17</td>
<td>4</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Upwork</td>
<td>Commitment to Nondiscrimination, Inclusion, and Respect</td>
<td>Separate</td>
<td>308</td>
<td>21.90</td>
<td>10</td>
<td>7</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Thumbtack</td>
<td>Non-Discrimination Policy</td>
<td>Separate</td>
<td>152</td>
<td>18.60</td>
<td>11</td>
<td>3</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Jobstack</td>
<td>Equal Employment Opportunity Policy</td>
<td>Separate</td>
<td>412</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Sittercity</td>
<td>Community Inclusion Policy</td>
<td>Separate</td>
<td>132</td>
<td>40.93</td>
<td>0</td>
<td>0</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Withlocals</td>
<td>Code of Conduct for Withlocals Guests</td>
<td>Combined</td>
<td>83</td>
<td>-</td>
<td>0</td>
<td>0</td>
<td>X</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>347.56</td>
</tr>
</tbody>
</table>

(table cont’d.)
4.5.7. Comparing domains

In terms of application domain, lodging apps in our dataset (except for Vrbo) have the highest quality NDPs (length = 700 words, readability = 40.5, types = 7, examples = 9). The asset-sharing app Turo has slightly higher numbers, however, it is the only app in its domain that has a NDP. Apps in the ride-sharing domain seem to have low quality NDPs in comparison to other domains (length = 107 words, readability = 9.4, types = 10, examples = 0). These results are surprising given that existing literature provided significant evidence of systematic bias affecting these apps [6, 11]. The same applies to delivery apps, however, these apps did not receive as much attention as ride-sharing apps in the digital discrimination literature [85]. While skill-based apps (length = 248 words, readability = 13.6, types = 10, examples = 2) are slightly better than ride-sharing apps, they still lag behind lodging apps. This is also surprising given that apps in this domain are known to have serious discrimination issues, such as bias against women and black workers, including lower hourly rates, lower ratings, and racially and sexually charged reviews [13, 14, 56, 8].

4.6. Discussion and Impact

Given the general shift in society towards more equality and prosperity, we anticipate that NDPs are going to become mandated by law in the near future. However, in the absence of a standardized format and the lack of regulations, drafting such policies remains a challenging and time-consuming task. To help overcome these challenges, the framework presented in this chapter provides developers with a systematic protocol for evaluating their policies based on their intrinsic characteristics and by comparing them to existing high-quality NDPs. This framework can also help developers to keep their NDPs in-check during software evolution. This can be particularly important for start-ups, where it can be financially infeasible to hire a third-party firm to take care of the policy as the system evolves and as we learn more about the problem.
Our work in this chapter bridges an important gap in the software maintenance and evolution research by focusing on non-code artifacts. Maintaining software policies is a prime example of adaptive maintenance tasks, where an artifact has to constantly change in order to adapt to external factors, such as changing regulations. In fact, such policies can be used to monitor the evolution of the system by monitoring changes to the NDP. Existing research suggests that important information about the system can be inferred from the modifications made to its privacy policy [112]. Furthermore, providing informative, comprehensive, and accessible NDPs can help users to make more informed decisions in the sharing economy market. In particular, users often find themselves having to choose from among hundreds of sharing economy platforms. The ability to make the right decisions in such a volatile market is critical for users to maximize their social and economic gains [2, 117].

In terms of results, our analysis shows that quality of NDPs varies among apps and application domains. Lodging apps seem to have the highest quality NDPs, while ride-sharing and skill-based apps do not provide informative NDPs. In terms of individual apps, the vehicle-sharing app Turo and the lodging app Airbnb provide the most comprehensive policies. Another observation is that apps do not mention in their NDPs the design strategies they use to mitigate discrimination. For instance, to control for bias in reviews, Airbnb rolled out a design change to ensure that hosts and guests can see the reviews only after both parties have submitted their reviews. According to Airbnb, “Both hosts and guests may worry that if they leave an honest review that includes praise and criticism, they might receive an unfairly critical review in response. To address this concern, reviews will be revealed to hosts and guests simultaneously” [70]. However, such a feature update is only mentioned in the blog maintained by Airbnb and is not highlighted in the NDP.

Our recommendation for developers drafting their own NDPs is to refer to apps’ with high quality policies (e.g., Airbnb and Turo) as good industry standards and to keep up with existing non-discrimination regulations. Furthermore, developers should always refer
to emerging research on digital discrimination. Such research constantly exposes problems of bias in sharing economy as well as suggests and evaluates mitigation strategies for these problems [10, 55, 54].

In terms of limitations, the main threat to the external validity of our study stems from the fact that only 108 popular sharing economy apps were considered in our analysis. However, as mentioned earlier, discrimination issues are more likely to manifest over these apps rather than smaller apps which typically target homogeneous populations of users. Furthermore, our search process utilized multiple search strategies and inclusion criteria to locate a representative sample. Generally speaking, the size of the dataset is aligned with datasets typically used in policy analysis research [111, 112, 113]. Another threat might stem from the fact that our evaluation of NDPs was conducted manually. Nonetheless, manual inspection of policies is a common practice in such kinds of studies. This threat can be mitigated by using a systematic review process and a well-defined review protocol with multiple judges. Furthermore, the majority of evaluation measures were quantitative in nature, therefore, subjectivity threats were minimized. Other concerns might be raised about the measures or the questions used in the evaluation protocol [112, 113, 106]. However, the majority of these measures were adapted from well-established protocols for evaluating privacy policies as well as existing regulations. These measures capture to a large extent the different aspects of NDPs in the sharing economy market. Finally, to summarize our findings in this chapter, we revisit our research questions:

- **RQ1:** *How prevalent are NDPs in the sharing economy market?* Our analysis of 108 sharing economy apps shows that NDPs are not common. Most apps either do not provide a NDP at all or provide a very brief and generic statement. Only a few apps maintain a separate NDP. Such policies appear under various names. This might negatively impact their discoverability and accessibility [111].

- **RQ2:** *Can the quality of existing NDPs be systematically evaluated?* Existing anti-discrimination regulations as well as protocols for evaluating the content
of software privacy policies can be adapted to NDPs. Specifically, NDPs can be evaluated based on a set of measures that can be extracted directly from the policy. These measures include quantitative metrics, such as the policy’s length and its readability as well as the number of examples and types of discrimination acknowledged in the policy, along with more qualitative measures, such as whether the policy describes any measures taken to mitigate discrimination and how cases of violation are reported and handled. While these measures capture all the aspects of NDPs, other, more complex, measures which go beyond the surface characteristics of policy text can be used.

- **RQ$_3$: How detailed and informative are existing NDPs?** Our analysis shows that the majority of NDPs in the sharing economy market are of low quality. Either they are very brief or do not provide sufficient information on what is considered discriminatory behavior or how that behavior is controlled for through the functional features of the app.

4.7. Conclusions

In this chapter, we presented a framework for evaluating NDPs of sharing economy apps. Our framework is based on an assessment protocol which uses a set of predefined measures to evaluate the quality of NDPs. Our results showed that most sharing economy platforms do not provide any form of NDPs. The results also showed that most of the NDPs are either brief, combined with other existing policies, or do not include essential information that is necessary to outline the app’s stance on discrimination. On average, apps in the lodging domain provide the most comprehensive policies, while apps in other domains still lag behind. Our work in this chapter aims to help software developers working with sharing economy apps to draft and maintain effective NDPs for their apps as well as help users to realize their rights to be treated fairly in one of the fastest growing software ecosystems in the world. Finally, our work in this chapter will be extended across two main directions:
• **Automation.** We will use text mining and modeling techniques to automatically learn the structure of NDPs, the main topics they discuss, and eventually generate an overall quality score for the policy. A fully automated prototype will be made publicly available to help app developers around the world draft high quality NDPs.

• **User studies.** Automated quality metrics, such as readability, can provide an indication of NDPs’ accessibility to the casual user. However, to enable a more objective assessment, user studies must be conducted. Such studies will involve recruiting large samples of the sharing economy users (providers and receivers) and using systematic questionnaires to assess their understanding of NDPs.
Chapter 5. Framework

In this chapter, we incorporate methods of requirements modeling and domain engineering to construct a conceptual framework for modeling discrimination concerns in the DSE market. Our objective is to represent such a complex domain phenomenon using simplified formal notations that software engineers, working in highly agile environments, can easily interpret and communicate, and effectively integrate into their DSE app design.

The comprehensibility of the proposed models is evaluated using an empirical study with 12 software professionals. We further conduct a case study at two software startups to examine the influence of our models on the design decisions of DSE app developers. The results show that our framework can facilitate a quick transition from complex domain knowledge to software requirements, enabling software engineers working in agile environments to effectively integrate discrimination mitigation strategies into their app design.

5.1. Introduction

Our SLR in Chapter 2 has revealed that the research on the design aspects of DSE software which enable a complex socio-technical phenomena such as digital discrimination to emerge online remains underdeveloped. This can be partially attributed to the lack of a systematic framework for consolidating existing scattered evidence into unified formal representations that software engineers can effectively interpret and communicate [118]. To bridge this gap, in this chapter, we propose a conceptual framework for modeling discrimination concerns in the DSE market.

A framework can be defined as a “reusable design (models and/or code) that can be refined (specialized) and extended to provide some portion of the overall functionality of many applications” [119]. Our framework is intended to help DSE developers to understand how the interactions between their functional features and user goals can facilitate bias and differential treatment of DSE users, and ultimately, deliver DSE design solutions that can promote equality and mitigate bias.
The proposed framework can be particularly important for developers operating in volatile and high-risk environments, such as software startups. Developers in software startups often find themselves under immense pressure to generate an initial set of requirements and deliver a working product [120]. Therefore, requirements in such contexts are typically generated using agile methodologies that are designed to enable software engineers to learn fast and act quickly (e.g., brainstorming and rapid prototyping) [121, 122]. The objective of our models is to increase the effectiveness of such methodologies by externalizing and representing complex domain knowledge through concise representations that can be quickly comprehended and effectively integrated into working prototypes. In the long run, the impact of the proposed research will extend to the entire population of DSE users, helping people in marginalized groups and resource-constrained communities to overcome key barriers to participation in DSE activities.

To construct our framework, we adapt an evidence-based approach. Evidence-Based Software Engineering (EBSE) aims to provide the means by which current best evidence from research can be integrated with practical experience and human values in the decision-making process regarding the development and maintenance of software [30]. In other words, decision making is grounded in the findings of research studies [87]. Kitchenham et al. [30] identified five main steps to conducting EBSE studies. These steps are: a) converting the need for information into answerable questions, b) tracking down the best evidence to answer the questions, c) critically appraising that evidence for its validity, impact, and applicability, d) integrating the critical appraisal with our software engineering expertise, and e) evaluating the effectiveness and efficiency of the process and seeking ways to improve it.

In the second chapter, we systematically synthesized existing interdisciplinary evidence on digital discrimination. In this chapter, we attempt to answer the following research questions:
• **RQ₁**: *Can evidence on digital discrimination be integrated into software domain models?*

• **RQ₂**: *Can such models be effectively understood and communicated by software professionals?*

• **RQ₃**: *Can such models influence app design decisions during requirements elicitation sessions?*

5.2. Modeling Framework

In information-intensive systems, knowledge externalization is often achieved via modeling [123]. Models provide a framework for explicitly describing abstract salient concepts in a specific domain and formally reasoning about these concepts in order to create new knowledge [124, 125, 126]. Several domain modeling techniques, such as Feature-Oriented Domain Analysis (FODA) [127] and Softgoal-Interdependency Graphs (SIGs) [128], have been proposed in the Software Engineering literature [66, 129]. Derivations of these models are commonly used in Software Product Lines (SPLs) to maximize systematic reuse of software assets and minimize production cost [130, 131]. In our framework, models are intended to capture the underlying feature-goal interactions of DSE platform that are related to digital discrimination. To achieve this goal, we adapt the following types of domain models:

- **Feature models (FM)**. These models are commonly used in requirements engineering analysis to show the core features of a family of applications in a specific domain along with their commonalities and variabilities. A domain in software engineering refers to any group of functionally-related software systems [127].

- **Feature-softgoal-interdependency graphs (F-SIG)**. These models can capture the complex interactions between the features and goals in the domain. Thus, helping software engineers to establish an understanding of their domain at an early stage of the software production process.
• Feature-goal-discrimination models (FGD). We introduce these hybrid models to integrate information from our systematic review into our domain models, showing the specific relationships between system features, user goals, and concerns of discrimination.

In this section, we describe the main entities of each of these models along with our model building procedures.

5.2.1. Feature Analysis and Modeling

Feature analysis is crucial to understand the basic feature interactions in the domain before the feature-goal interactions can be captured [132]. Our domain of analysis consists of the DSE platforms that are commonly investigated in related literature (Fig. 2.1), including the ridesharing services Uber and Lyft, the lodging service Airbnb, and the freelancing services TaskRabbit, Upwork, and Fiverr.

• Feature Matrix

While several automated feature analysis solutions have been proposed in the literature [133], the process is still largely manual, relying on a qualitative analysis of existing feature documentation [134]. To extract the main features of our domain, we follow Nedić et al. [135]’s principles for Feature Model (FM) engineering. These principles cover the different phases of feature modeling, from planning over model construction, to model maintenance and evolution. Following the principle of “rely on domain knowledge and existing artifacts to construct the feature model”, we extracted the feature information of our subject platforms from three main sources: their marketing websites, descriptions on app stores, and direct execution. We observed that marketing websites were the most informative source of feature information. In fact, this was not surprising given that app store descriptions are known to be short and uninformative [136]. Marketing websites, on the other hand, provide detailed descriptions of features, often indexed into organized views.
Extracted features were then organized into a detailed feature matrix. Feature matrices list the main features of a group of systems in a certain domain [137]. Following the principle of “features at higher levels in the hierarchy should be more abstract”, a requirements analysis session was held to identify features that are common to most systems in our domain (top of their hierarchies). In particular, we mirrored the open coding process we followed for synthesizing evidence in Chapter 2. Three software professionals (average four years of industrial experience in software design and engineering) went through the list of features individually and categorized them based on their specific functionality. Feature categories started quickly emerging as most DSE systems implement almost the same set of high-level functionality. Individual classifications where then compiled into a list of categories that can be described as follows:

**P2P Connection.** At a very abstract level, the core feature of DSE platforms is to establish a P2P connection between service providers and receivers, such as riders and drivers, renters and hosts, and business owners and freelancers. This feature is common to all DSE platforms. Varieties of this feature include fully automated matching (e.g., matching based on geolocation in Uber) or semi-automated matching. For example, matching in TaskRabbit is first automatic based on the required skill and then manual based on user preferences of individual **taskers**.

**Payment.** This bundle of features handles money transactions between users and the platform, including paying for the service, refunds, credit, and fees. Some platforms provide more payment options, including tipping, pay-with-cash, payment plans, such as the payment plans of Upwork, and share payment, such as the split-fare and group-pay features of Uber.

**Transaction management.** Almost all DSE platforms provide features to manage a transaction after a connection is established. Users can confirm (e.g., book the place) or cancel (e.g., cancel ride) the transaction. Some platforms elaborate on this feature by providing a transaction tracking functionality, such as the status of the task being
performed, or how far the Uber driver is. Features for managing transactions’ history and favorite transactions are also sometimes provided. Some platforms go a step further to enable conflict resolution, such as the dispute resolution process of TaskRabbit.

**Reputation systems.** Ratings and reviews, or reputation systems, are commonly used in DSE platforms to establish trust [84, 42]. Users often resort to the provided rating system to rank their experiences with other users (providers or receivers). Varieties of this feature include rating on a 1-5 star scale, thumbs up or thumbs down, leaving a text review, or reporting users (a rude rider or driver, a dirty rental place, or a poorly skilled freelancer). Furthermore, some platforms enable users to boost their reputation by awarding badges to users who have longer service records and consistently provide the highest level of service and professionalism, such as Superhost in Airbnb (Fig. 2.3) or Elite Tasker in TaskRabbit.

**Direct messaging.** Direct messaging is an essential feature that connects DSE service providers and receivers. Some platforms, such as TaskRabbit and Airbnb, allow users to communicate through the secure in-app messaging feature to enhance safety.

**User profile.** DSE platforms typically maintain some sort of a profile for their users, including their names, emails, and pictures. This feature is accompanied with features to manage such information (e.g., change picture, location, preferences, etc.). Varieties of this feature include more elaborate profiles. For instance, in freelancing apps, users are expected to list their set of skills as a part of their profile. Furthermore, some systems provide external verification mechanisms to verify the identity of their users. Fig. 2.3 shows an Airbnb host’s profile with a verified identity.

**Safety.** Safety has appeared among our list of goals. We find it appropriate to use it to label the bundle of features related to safety. Specialized safety measures take different forms. For instance, in response to recent incidents of abduction, Uber rolled out a new set of features to enable users to text a SOS message along with trip information to 911 directly from the app. Lyft has a similar feature to privately request help right from the app. Additional features include sharing trip details with a friend and using a PIN number.
to verify the identity of the driver and the passenger. Airbnb implements a risk scoring feature where a reservation is scored for risk before it is confirmed, scam prevention, account protection, and secure payments are also implemented in most platforms.

**Accessibility.** Similar to safety, the goal of accessibility is also enforced through a bundle of features. For instance, Uber provides a specialized service in selected cities known as Wheelchair-Accessible Vehicles (WAVs) to accommodate people with wheelchairs and service dogs. Airbnb enables hosts to highlight features (accessibility profiles) of their homes that make them accessible to guests who use wheelchairs, canes, or other mobility aids. Lyft offers a Hard-of-Hearing driver feature which sends passengers instructions to use text messages rather than call their drivers.

- **Feature Modeling (FM)**

Features matrices are best represented through Feature Models (FMs) [138]. FMs play a *de-facto* role in understanding the functional features of a specific domain, facilitating tasks such as model driven development [139] and feature oriented programming [140]. In this chapter, due to space limitations, we generate a sample model for the payment feature only. This model, shown in Fig. 5.1, shows that all platforms in our domain provide some sort of a payment method, a variety of sub-features of this feature include providing multiple payment options. Similar models can be generated for other features using the supplemented feature matrix.

### 5.2.2. Feature-Goal Modeling (F-SIGs)

Our review exposed four user goals that are frequently related to discrimination. To represent these goals, along with their interactions with features, we adapt Feature-Softgoal Interdependency Graphs (F-SIGs) [141]. In F-SIGs, functional features are represented using a rectangle shape and softgoals are represented using a cloud shape. The edges of the graph represent the interrelationships among the softgoals and features. These relationships are represented using arrows accompanied with plus and minus signs to indicate the type
of impact among softgoals and features: (-) for hurts, (--) for breaks, (+) for helps, and (++) for makes.

We start our modeling construction process by connecting our goals. Users do not participate in DSE activities without a minimum level of trust in the platform. Safety is a must to achieve trust. Even if service providers or receivers are trustworthy (high ratings or only have positive reviews), a transaction will not take place if there is any perceived risk associated with the process. Finally, the accessibility goal enhances inclusion for people with disabilities. To show the relationships between features and goals, a (++makes) is assigned for features that are common to all platforms. For instance, profile information and reporting are implemented by all platforms as a means of safety, while reputation systems are provided by all platforms to establish (++makes) trust. Features that are implemented by only some platforms are assigned a (+helps) relationship to goals. The generated feature-goal model for our domain is shown in Fig. 5.2. It is important to point out that this model is partial. It only captures the interdependencies between our domain features and goals that have been identified in our literature review. Other goals such as usability, reliability, and ecological sustainability, while important in DSE, are irrelevant to the problem of discrimination.
5.2.3. Feature-Goal-Discrimination Models (FGDs)

The main goal of our analysis is to integrate concerns of digital discrimination into existing domain models using notation that software engineers can easily interpret. This goal is aligned with Kitchenham et al.'s EBSE guideline of integrating extracted evidence with software engineering expertise [30]. Generally speaking, a user concern, or an anti-goal, can be defined as any functional or non-functional behavior of the system that might negatively impact its end-users experience or their overall well-being [142]. Software user concerns can range from mental and physical to economic, political, or even cultural concerns. In our analysis, concerns are specific to acts of discrimination that are facilitated by DSE platforms (Fig. 2.2). Our review of existing literature uncovered significant links between these concerns, user profiles, and reputation systems. The main research question at this phase is, how can such information be integrated into our domain models?

Conventional domain modeling techniques (e.g. FMs and F-SIGs) do not provide an effective mechanism for integrating concerns of discrimination. To work around this limitation, we slightly alter the semantics of F-SIG notation. Specifically, in our previous work [10], we used shaded clouds to represent concerns of digital discrimination. Further-
Figure 5.3. A feature-goal-concern diagram of the profile feature (⊖ ⇒ mitigates, ⊕ ⇒ triggers).

more, we used an arrow with a minus circle head (⊖) to indicate a mitigates relationship (e.g., a specific feature mitigates a specific type of discrimination) and an arrow with a plus-circle head (⊕) to indicate a triggers relationship (e.g., a feature leads to a type of discrimination). This notation is chosen to be simple. Simplicity of notation is critical for domain models’ scalability, extensibility, and understandability [135]. To generate our models, we establish our model relationships based on the evidence synthesized from the literature. For clarity purposes, we generate two models: one for reputation systems and one for user profiles.

Fig. 5.3 and Fig. 5.4 show the models generated for the profile and the reputation features respectively. Both diagrams show the discrimination concerns stemming from these features and their sub-features along with the strategies suggested in the literature to mitigate these concerns. Discrimination concerns are placed at the bottom and grouped using dashed boxes to enhance the visual appearance of the models.
5.3. Model Evaluation

In the first phase of our analysis, we used an enhanced modeling notation to integrate evidence on digital discrimination into DSE domain models. The question at this phase of our analysis is: **RQ₂ Can such models be effectively understood and communicated by software professionals?**. Model evaluation can be a challenging task. An effective model is a model that is correct (captures the actual reality of the domain) and can be effectively understood and communicated. While formal structural metrics calculated over the model graph (e.g., number of nodes, leaves, and their degree of connectivity) [143] can give an indication of its complexity, empirical evaluation remains a must to evaluate a construct as subtle and complex as comprehensibility [144, 145]. In this section, we empirically evaluate our models’ comprehensibility. Our objective is to assess the value of our generated models to software engineers as project communication artifacts [144, 143].
5.3.1. Method

Our empirical evaluation follows the main guidelines proposed by Aranda et al. [144] for model evaluation in requirements engineering. In particular, the authors proposed four model comprehensibility variables:

- **Correctness of understanding.** The degree to which a study participant can correctly answer questions about the domain captured in the model.

- **Time.** The time required for a study participant to answer questions based on the model.

- **Confidence.** The subjective measure of the perceived confidence of a study participant in their answers.

- **Difficulty.** The subjective judgment that a study participant displays regarding the ease to obtain information from the model.

In the literature, such variables are often measured using questionnaires, including a mixture of close and open-ended questions [146, 144, 147]. To evaluate our models, we design a questionnaire with a set of seven questions. These questions are shown in Table 5.1. In general, our questions can be divided into three main categories. \( Q_1, Q_2, \) and \( Q_3 \) are intended to measure the correctness of understanding by asking questions about the specific interdependencies in the model. \( Q_4 \) and \( Q_5 \) are intended to assess the perceived difficulty of the model. \( Q_6 \) is a measure of the perceived confidence in the answers provided, and \( Q_7 \) is to control for the interaction of history and treatment, in other words, determine whether the answers originated from previous knowledge of the problem rather than the models.

5.3.2. Subjects and Procedure

Our set of study participants (subjects) consisted of 12 developers sampled from four software startups through personal connections (convenience sampling). All participants
Table 5.1. The model comprehensibility questionnaire used in our empirical evaluation.

<table>
<thead>
<tr>
<th>Code</th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q₁</td>
<td>How would you describe the relationship between trust cues and discrimination?</td>
<td>Open-ended</td>
</tr>
<tr>
<td>Q₂</td>
<td>How would you control for discrimination based on pictures?</td>
<td>Open-ended</td>
</tr>
<tr>
<td>Q₃</td>
<td>Would entirely removing reputation systems be a reasonable way to eliminate the problem of discrimination?</td>
<td>Open-ended</td>
</tr>
<tr>
<td>Q₄</td>
<td>How hard were the models for you to interpret?</td>
<td>Very easy, Easy, Hard, Very hard</td>
</tr>
<tr>
<td>Q₅</td>
<td>Which part(s) of the models was/were the hardest to interpret?</td>
<td>Open-ended</td>
</tr>
<tr>
<td>Q₆</td>
<td>How confident do you feel that you can communicate this knowledge with other software professionals?</td>
<td>Not confident, Somewhat confident, Confident, Very confident</td>
</tr>
<tr>
<td>Q₇</td>
<td>Do you have any former knowledge of digital discrimination?</td>
<td>I had no knowledge at all, I had some knowledge, I had thorough knowledge</td>
</tr>
</tbody>
</table>

are Computer Science/Software Engineering graduates with an average of four years of experience in professional software development. They all assume multiple rules in their startups, from requirements engineering, to system design and architecture, implementation, and testing. In addition to their core projects, the four startups also work on multiple projects, including mobile applications, web services, and API development. Only two participants reported experience in building DSE systems, including a system for hiring tech freelancers. All participants reported using DSE platforms, including Uber, Lyft, Airbnb, Wag!, and DoorDash. Each subject was compensated $24 as an incentive for their participation in the study. The procedure of our experiment can be described as follows:

- We met with each group of study participants from each startup separately. The meetings were held in person in a conference room setup. One researcher attended each session. The subjects were informed that the goal of the study was to help software engineers understand the problem of digital discrimination in DSE. To minimize
threats to validity, we did not inform the subjects of the fact that the models were
generated by us.

• Each subject was then presented with our assignment sheet, which included an intro-
duction to the problem, the goal of the study, the models in Fig. 5.3 and Fig. 5.4, a
notation key, and a description of the different features and goals in the model. The
description did not include any information about how these features were related to
goals or discrimination concerns other than what was displayed in the models.

• The subjects were given one hour to study the proposed models. They were allowed
to ask questions during this session. They were informed that they did not have to
spend the full hour studying the assignment. They just had to inform the researcher
whenever they were ready to answer the questionnaire.

• We then presented our questionnaire to each subject. No time constraint was en-
forced. Our subjects were not allowed to ask questions anymore.

• The responses, along with the time it took for each subject to study the assignment
and answer the questions were collected and the data was analyzed.

5.3.3. Understandability Analysis

To analyze and report the answers of the open-ended questions, we follow a grounded
theory procedure. In particular, we use open coding and memoing to categorize responses
to the open-ended questions and then selective coding to report the results [69]. The
open coding process involved all authors going through questions Q1, Q2, Q3, and Q5,
identifying salient response categories as they appeared in the text. The process was
carried out over an hour-long session. The different categories were then consolidated into
multiple response categories as follows:

Q1: How would you describe the relation between trust cues and discrimi-
nation? Five (N=5) subjects provided the right answer with no further elaboration (trust
cues mitigate racism in the reviews). Five (N=5) subjects elaborated by explaining that according to the model, when people see a badge next to a service provider name they are less likely to leave a racially-charged comment. Two (N=2) subjects acknowledged the mitigates relation, but demanded more information on why that was the case.

Q2: How would you control for discrimination based on pictures? Ten (N=10) out of our subjects provided a correct answer based on the model, indicating that they would delay exposure to pictures until the transaction is finished and encourage users to use an asset-based picture. Two subjects (N=2) indicated that they were not sure what kind of asset picture they would use, for example to promote a skill asset such as coding. They implied that they would delay exposure to information but not use asset pictures. One subject (N=1) raised concerns about the impact of using an asset picture rather than a profile picture on the safety of users.

Q3 Would entirely removing reputation systems eliminate the problem of discrimination? Two subjects (N=2) answered with yes. If discrimination gets out of control, abandoning this feature all together seems like a good solution. However, seven subjects (N=7) answered with no, trust cannot be established without some sort of a reputation mechanism. Several subjects (N=5) emphasized the importance of having some sort of a reputation system in the absence of an alternative objective mechanism for building trust. Two other subjects (N=2) viewed the problem as a pros vs. cons. One subject stated that “I would say they should be kept as they are essential to do business with strangers online, but if discrimination gets out of hand, then maybe they should be removed.”. One last subject (N=1) was not able to correctly answer the question, writing “I am not quite sure.”

5.3.4. Difficulty and Confidence

The results of the close-ended questions are shown in Table. 5.2. The results show that, in terms of easiness (Q4), the majority of our subjects found the models to be easy while two subjects found them to be hard, no subjects indicated that the models were either
Table 5.2. A summary of our questionnaire’s participants’ answers.

<table>
<thead>
<tr>
<th>Subj.</th>
<th>Startup</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>C1</td>
<td>Easy</td>
<td>Confusing notation</td>
<td>Confident</td>
<td>No knowledge</td>
</tr>
<tr>
<td>S2</td>
<td>C1</td>
<td>Easy</td>
<td>Confusing notation</td>
<td>Confident</td>
<td>No knowledge</td>
</tr>
<tr>
<td>S3</td>
<td>C1</td>
<td>Hard</td>
<td>Information sparsity</td>
<td>Somewhat confident</td>
<td>No knowledge</td>
</tr>
<tr>
<td>S4</td>
<td>C2</td>
<td>Easy</td>
<td>Confusing notation</td>
<td>Confident</td>
<td>No knowledge</td>
</tr>
<tr>
<td>S5</td>
<td>C2</td>
<td>Easy</td>
<td>Confusing notation</td>
<td>Confident</td>
<td>No knowledge</td>
</tr>
<tr>
<td>S6</td>
<td>C2</td>
<td>Easy</td>
<td>Information sparsity</td>
<td>Somewhat confident</td>
<td>No knowledge</td>
</tr>
<tr>
<td>S7</td>
<td>C3</td>
<td>Easy</td>
<td>Confusing notation</td>
<td>Somewhat confident</td>
<td>Some knowledge</td>
</tr>
<tr>
<td>S8</td>
<td>C3</td>
<td>Hard</td>
<td>N/A</td>
<td>Not confident</td>
<td>No knowledge</td>
</tr>
<tr>
<td>S9</td>
<td>C4</td>
<td>Easy</td>
<td>N/A</td>
<td>Confident</td>
<td>No knowledge</td>
</tr>
<tr>
<td>S10</td>
<td>C4</td>
<td>Easy</td>
<td>Confusing notation</td>
<td>Confident</td>
<td>No knowledge</td>
</tr>
<tr>
<td>S11</td>
<td>C4</td>
<td>Easy</td>
<td>N/A</td>
<td>Confident</td>
<td>No knowledge</td>
</tr>
<tr>
<td>S12</td>
<td>C4</td>
<td>Easy</td>
<td>Information sparsity</td>
<td>Somewhat confident</td>
<td>No knowledge</td>
</tr>
</tbody>
</table>

very easy or very hard to interpret. To elaborate more on this question, we analyzed the answers to Q5: Which part(s) of the models was/were the hardest to interpret? The results show that our subjects faced two problems analyzing our models:

- **Information sparsity.** Three (N=3) out of our 12 subjects implied that information sparsity was the main issue in interpreting the models. For example, S₃ demanded more information on how bias correction can be achieved, claiming that just showing ”Bias correction” in the model does not provide enough information on how exactly to implement this feature. This limitation suggests that more supporting information should be provided with models.

- **Confusing notation.** Six subjects (N=6) indicated that the notation was in some cases confusing. One subject elaborated that they did not understand the difference between the mitigates/trigger and mandatory/optional notation and another subject implied that they did not understand user goals to begin with. This was actually
expected as developers unfamiliar with modeling notation can find it hard to interpret such notations at first. We expect that this effect will diminish as developers become more familiar with the model semantics.

In terms of confidence ($Q_6$), 11 out of 12 subjects indicated that they were either confident or somewhat confident in communicating their knowledge about discrimination to others. One subject indicated that they were not confident. In their answer to $Q_4$, they wrote that they were “not good with models and graphs in general.”

5.3.5. Time Analysis

Our results in terms of time (shown in Fig. 5.5) show that our subjects took on average 41 minutes to go through the models and the keys and on average 21 minutes to answer the questions in our questionnaire. We conclude from our study that our models were successful in communicating design information about the problem. In summary, the majority of our subjects found the models to be easy to interpret and they were able to provide reasonable answers based on the models in a reasonable amount of time, giving a strong indication of the effectiveness of the models in communicating domain knowledge. These results are encouraging given that the overwhelming majority of our subjects indicated that they had no previous knowledge of the problem.
5.4. Extrinsic Evaluation

The objective of our extrinsic evaluation is to assess, through experience, whether our models would influence DSE app design. Our research question is (RQ₃) *Can our models influence app design decisions during requirements elicitation sessions?* We conduct our extrinsic evaluation at two software startups. Startups are different from traditional established companies in the sense that they have to immediately and accurately identify and implement a product that delivers an actual consumer value. Such a product is commonly known as the Minimum Viable Product (MVP) [148, 149]. The requirements for MVPs tend to be market-driven, generated through agile requirement elicitation methodologies [121, 122]. Our expectation is that our proposed models will help startups, operating under significant time and market pressure and with little development history, to gain a quick and comprehensive understanding of their domain of operation, and thus, serve as an integral part of their MVP design process.

5.4.1. Method

We selected two startups from the four startups that participated in our study to conduct our extrinsic evaluation. The first startup has three developers and the second has four. All seven developers have participated in our model evaluation procedure. Our extrinsic evaluation takes the form of a case study of requirements elicitation of an DSE app in an agile environment. Our subject app is a P2P app for hiring hair stylists. A user can subscribe either as a stylist or as a regular customer. The app then matches customers with stylists based on geolocation, price range, gender, and the hair job required (haircut, dye, style, etc.). The transaction is expected to take place at the customer's premise. Both the stylist and the customer can rate their experience afterwards. Payments are facilitated through the app.

With this generic product statement in mind, the team of software engineers from each startup were then asked to come up with their main system features, focusing on what would go into their MVP. One session took place at the startup while the other session
was virtual. The researchers were not present during both sessions to avoid influencing the design in any way. Both teams were told to use whatever notations (textual user stories or visual diagrams) they wanted. An hour and a half was set as the time for the session. Both startups reported that their brainstorming sessions typically last between 60 and 90 minutes. Each subject from both startup was paid $36 to participate. Both sessions were conducted on the weekend to avoid interfering with our subjects’ work schedules. A brief interview was then held after the session. In what follows, we analyze the design decisions of each start-up and the outcome of our interviews.

5.4.2. Results

The team from the first startup used a white board to sketch out their design. The sketch is shown in Fig 5.6. The design takes the form of a basic usecase diagram. Two main actors were identified, the stylist and the customer, and multiple use-cases were considered. According to this design, both actors have access to the Profile, Matching, and Review features. The stylist has access to the Schedule Work feature which enables them to announce the times they are available for service. The customer has access to the Request Service feature through which they can look for a stylist. In addition, customers have access to the Payment feature through which they can pay their stylists. The team also utilized the $<$extends$>$ and $<$includes$>$ relations of usecase diagrams, for example, reviews include a mutual review mechanism while matching can be extended by auto-matching.

Fig. 5.7 shows the diagram produced by the second team. This team used a feature model to show the main features of their app. The team used an online drawing tool as the meeting was virtual. The diagram has Profile, Search and Match, Payment, Reputation, and Work Settings features. The diagram also includes sub features of the main features. For example, payments can be made either through credit or debit cards or PayPal and a stylist can set their work hours and their rates through the Work Settings feature. In general, both diagrams by both startups included most of the core features of DSE apps.
To answer RQ$_3$, we followed up with a semi-structured interview, where we asked each group of software engineers from each company three open-ended questions about their design. The interviews were recorded and the audio was then analyzed and coded by the authors. Key points from the discussion are reported next:

- **Did you consider any anti-discrimination features in your design?** The team from the first startup added three features to counter discrimination, Mutual Reviews, Social Verification, and Automated Matching. They indicated that mutual reviews would help to prevent retaliation in reviews. The team also indicated that socially verifying users can be used as a trust signal. One developer from the startup
elaborated by "I tend to trust users whom I can see on social media. This makes them look more like people rather than just service providers." The team also indicated that providing an optional fully automated matching feature (extension of the original feature) would prevent people from making biased decisions. Users who utilize this feature can be rewarded.

The team from the second startup added Withholding Information, Badges, and Structured Reviews as features to counter discrimination. Their justification was that most people who would be using this app would not be necessarily interested in a particular stylist, thus, no pictures or names are needed ahead of time. As for stylists, they should be able provide the service that match their skills without discrimination. Personal information would only be revealed after the transaction is confirmed. In case of cancellation without a reasonable justification, the app would penalize the canceling party. The team also indicated that they would be using a structured reputation mechanism for customers to rate the skills, professionalism, friendliness, and prices of the stylist. No text reviews would be available. Stylists on the other hand, can rate customers on a scale from 1 to 5 without text reviews as well. A system architect from the startup stated that, "reviews are always going to be problematic, especially when you are rating people." Badges will be used as an additional signal to bridge the trust gap that might result from not allowing textual reviews.

- **Were there any features that you did not consider at all, or maybe will consider, but in the future?** The team from the first startup indicated that they would not withhold pictures or names. They elaborated by implying that they had to keep personal information for safety purposes, given that the app facilitates a very personal transaction. The team also indicated that they would consider adjusting ratings, but that would be a low priority feature as large amounts of data (user ratings and reviews) is first needed. The lead software architect of the team indicated that, "I
imagine this feature would need lots of data and some sort of a data science component to be implemented. We would like first to see evidence of bias in the reviews before we can consider the feature”.

The team from the second startup indicated that adjusting reviews after the fact would be a very controversial feature. They also added that they might implement a trust cue system or social verification features, but this feature would be added at later stages depending on the number of users. The team also stated that they would consider a feature for analyzing frequent cancellations to detect if they were associated with discrimination and then act accordingly.

• **Did the provided models help in any way during the brainstorming session?** The team from the first startup indicated that our models helped them to frame their thoughts and easily integrate anti-discrimination features into their design. According to their lead architect, “usually we don’t exactly stick to a specific diagram, instead you see multiple diagrams covering different bits and pieces of the project. Diagrams which show exactly how to integrate specific features to counter problems as complex as discrimination are always welcomed.”

The team from the second startup indicated that the models helped them in constraint identification during their brainstorming session. According to one of their software engineers, “having access to such models during the initial phases of problem formulation and MVP requirements analysis can save so much time at later stages.” Another engineer elaborated by, “we were actually not aware of the problem at all. It is easy to get blinded by the core features of a new app and forget about other important aspects of the problem.”

In summary, our extrinsic evaluation shows that exposing software engineers to our models has influenced their design decisions and the way they thought about some of their features. This was reflected in the set of features that were included in the design.
The models enabled our subjects to think through multiple alternative app designs before settling down on the final MVP.

5.5. Discussion and Impact

The research on digital discrimination has gained a significant momentum over the past four years. This can be attributed to the unprecedented widespread use of DSE systems and the general shift in society towards more equality and prosperity. As more research is conducted, it becomes harder for software engineers to keep up with this growing body of research. To address this limitation, our proposed framework is intended to systematically synthesize existing evidence on digital discrimination and present the results using notations that software engineers can comprehend, communicate, and eventually integrate into their working systems. In their 2016 Harvard Business Review piece, Fisman and Luca emphasized the need for such a framework to facilitate thinking through the available design choices and their implications on discrimination [118]. To that extent, our framework is intended to act as a vehicle for facilitating a quick transition from domain knowledge to requirements specifications. Through our proposed models, system designers can get insights into the complex interaction of features that could trigger or mitigate discrimination in their operational environment. Such information can be used to redirect effort in agile environments where developers do not have the time to research complex domain phenomena between product cycles.

In terms of general impact, the framework presented in this chapter can be reused for other types of evidence-based software design problems. In particular, user concerns in the mobile app market extend over a broad range of personal and societal issues, impacting our mental and physical health, privacy and security, cultural norms, economic status, and overall social structure. Due to their interdisciplinary nature, evidence on these concerns is typically published in a broad range of venues. Our framework can be adapted to consolidate such evidence in any domain, helping software designers and requirements engineers to address their concerns in their domain at an early stage of the process. This can be
extremely critical for app survival as recent evidence has shown that the failure to address user concerns not only compromises users’ experience, but can also have catastrophic consequences for app success. In fact, apps that do not adequately address their concerns are often deemed untrustworthy, unhelpful, or even abandoned by users [142].

Finally, we revisit our research questions in this chapter:

- **RQ₁** *Can existing evidence on digital discrimination be integrated into software domain models?* Extracted evidence on digital discrimination can be integrated into existing domain models given that the right notation is used. Generated models should explicitly capture the interdependencies (synergy and trade-offs) between discrimination concerns, user goals, and features of DSE platforms in order to provide a full picture of the domain.

- **RQ₂** *Can such models be effectively understood and communicated by software professionals?* Our empirical evaluation showed that our discrimination-aware domain models can be a powerful tool for achieving knowledge externalization and transfer. During our empirical investigation, our study participants were able to answer technical questions about the domain as well as communicate such knowledge effectively and with a reasonable level of confidence. These results encourage us to utilize our framework in other domains where human interaction with software systems can result in unwanted side-effects.

- **RQ₃** *Can such models influence app design decisions during requirements elicitation sessions?* Our extrinsic evaluation with two software startups revealed that providing our models to software engineers during requirements elicitation sessions can influence the design of the MVP. According to our subjects, our models provided them with a framework to think through multiple design alternatives as well as to help them in the constraint discovery process. Studies of requirements practices in startups revealed that, due to limited resources, software teams do not
spend much time learning about the domain [120]. Our models can help to overcome this problem by providing developers with a concise and abstract representation of complex domain phenomena using a simple visual notation that they can effectively comprehend and integrate into their working systems.

5.6. Limitations and Threats to Validity

Internal validity threats might stem from our model-generation procedure. In particular, different interpretations of existing evidence might lead to slightly different models. We attempted to mitigate these threats by only considering high quality evidence in our models and by using an enhanced notation that can show the general direction and type of the relation (mitigates and triggers) between model entities.

A main threat to the external validity of our study stems from the fact that only a few popular DSE platforms were considered in our analysis. Discrimination issues are more likely to manifest over these platforms rather than smaller platforms which typically target homogeneous populations of users, consequently, popular platforms receive significantly more attention in the literature. Furthermore, selecting mature platforms gives smaller platforms a chance to learn from the mistakes of the big players in the market [150].

Other external threats might also originate from our evaluation procedures. In both phases of our evaluation (survey and case study), software startups were used as our target population. However, startups are very active in the DSE domain. In fact, some of the major DSE platforms in today’s market, such as Turo, Uber, and DoorDash, were convinced at startups [151]. Furthermore, unlike established companies, startups are often under constant pressure to deliver with less time devoted to focus on learning and gathering information about the domain. This makes them ideal environments to run our analysis. Other threats might originate from the fact that our case study involved only two startups working on an MVP for a hypothetical DSE system. This might limit the generalizability of our results beyond our case study.
Finally, a construct validity threat might be raised about the reliability of the questionnaire used to evaluate the understandably of our proposed models. Several other survey methods such as direct interviews could be used. To mitigate this threat, our questionnaire included a mixture of open-ended and closed-ended questions, which covered several aspects of our models. While more questions could have been asked, we feel that the number of questions in our questionnaire was sufficient enough to examine the knowledge of our participants, yet avoid any fatigue issues. We further acknowledge the fact that different groups of participants might generate different outcomes. To mitigate this threat, our pool of subjects included software professionals who were selected from four different software startups. This helped to increase the validity of our results as well as eliminate several other threats that are often associated with other subject populations, such as students.

5.7. Conclusion and Future Work

In this chapter, we proposed a new framework for modeling discrimination concerns in the Sharing Economy. We performed a systematic feature-goal analysis over the set of platforms identified during the first phase of our analysis and represented the results using a set of feature and feature-goal domain models. We then proposed a new notation for integrating discrimination concerns into our domain models.

The generated models were empirically evaluated using 12 professional software engineers sampled from multiple software startups. Our results showed that our models can be easily comprehended and communicated by software engineers. We further conducted a case study at two different software startups to assess the influence of our models on DSE app design. The results showed that our models can facilitate an effective transition from tacit domain knowledge to requirements specifications.
Chapter 6. Conclusions

In this dissertation, we studied the problem of digital discrimination in the Sharing Economy market from a software engineering point of view. We first conducted a systematic literature review of 58 interdisciplinary primary studies to synthesize and categorize existing evidence on digital discrimination. Our results showed that sexism and racism were the most popular forms of discrimination, mainly affecting popular Sharing Economy platforms, such as Uber and Airbnb. The results also showed that these forms of discrimination were often enabled by features of profile pictures (e.g., pictures and names) and reputations systems (e.g., ratings and reviews). Several design strategies have been proposed to mitigate these concerns.

In Chapter 2, we systematically synthesized and analyzed existing evidence on digital discrimination to identify discrimination concerns that affect DSE platforms. We also identified functional features that trigger discrimination, as well as mitigation strategies that were proposed in the literature. Our results showed that racism and sexism are among the most common types that affect ridesharing, lodging, and freelancing domains. In terms of features, our SLR showed that user profiles and reputation systems are the main enablers of digital discrimination in DSE. A variety of strategies to reduce discrimination were proposed in the literature, including withholding information, asset-based profile pictures, and bias correction.

In Chapter 3, we quantitatively and qualitatively analyzed a large-scale of users feedback scarped from the Twitter feeds of eight popular Sharing Economy platforms. Our results showed that even though Twitter is commonly used for reporting issues of bias in the Sharing Economy, most tweets were naturally brief and did not convey enough information about the discrimination incidents encountered and reported by either service providers or receivers.

In Chapter 4, we proposed a systematic protocol for drafting and evaluating non-discrimination policies in the Sharing Economy market. Our results showed that the ma-
jority of Sharing Economy platforms either did not provide a non-discrimination policy, or provided a very brief one that did not include essential information to outline the app’s stance on discrimination. The analysis also showed that existing policies were often combined with other usage policies. Apps such as Airbnb and Turo provided the best quality policies while other popular apps, such as Uber and TaskRabbit, only provided a brief statement on equality.

In Chapter 5, we devised a framework for modeling discrimination concerns in the Sharing Economy. Our framework utilized feature and domain models to capture interdependency relations between the different types of bias affecting sharing economy platforms and the functional features and user goals of these platforms. The proposed models were empirically evaluated using an experiment with 12 software developers sampled from four different software start-ups. Our results showed that domain models can be a powerful tool for representing complex discrimination concerns using simplified notations that developers can understand, communicate, and utilize as an integral part of their software design process.
Works Cited


Vita

Miroslav Tushev, born in Ryazan, Russia, received his Specialist degree from The Russian Academy of National Economy and Public Administration (RANEPA) in 2014 and his Master’s degree from Louisiana State University in 2018. As his interest in computer science grew, he decided to stay at Louisiana State University and pursue his Ph.D. Upon completion of his Ph.D., he will be entering the software industry as a data scientist.