2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)

Implicit Gender Biases in Professional Software Development: An Empirical Study

Yi Wang[1] & David Redmiles[2]
[1]*Department of Software Engineering, Rochester Institute of Technology*
[2]*Department of Informatics, University of California, Irvine*
[1]yi.wang@rit.edu, [2]redmiles@ics.uci.edu

***Abstract*—It has been well-known that the software development profession lacks gender diversity, particularly in the technical leadership positions. Researchers and practitioners have spent tremendous efforts on identifying the problems and finding solutions. However, most of the existing software engineering literature focuses on the explicit gender biases but ignores implicit gender biases. To fill this gap, the study sought to empirically investigate whether professional software engineers hold implicit gender biases related to women in the software development profession, and examine whether these implicit biases predict discriminatory decision-making. Using data from 142 professional software engineers in seven organizations, our study yields a rich set of concerning findings. First, we find that implicit biases were pervasive–both male and female software engineers implicitly associated software development professions, particular technical leadership roles, with men, not women, and also associated women with the home and family. Besides, people often cannot resist their implicit gender biases and make decisions in gender-neutral ways while they do well in resisting their explicit gender biases.**

***Index Terms*—implicit gender bias, gender diversity in SE, decision-making, implicit association test (IAT), empirical study**

I. Introduction

Lack of gender diversity in the software development profession, particularly the leadership professions such as senior software architect, is a well-known and well-documented problem [1]–[5]. According to a recent labor market census, in 2015, female developers only account for 21% of the whole software development workforce and earn $19,514 less than their male peers [6] on average. Researchers in areas such as SE, education, management, etc., have spent tremendous efforts to solve the problem by creating new opportunities and pathways [7]–[11], setting code of conduct in workplaces and professional organizations [12]–[14], and so on. Policy-makers in many software development organizations have launched formal programs to fight against gender discrimination. We have witnessed a devoted office of diversity and inclusion is being established in many large organizations, e.g., Google[1], IBM[2], and so on.

These efforts are highly successful in reducing explicit gender-based biases and discrimination towards women in the software development profession. However, only addressing explicit gender biases is not enough. Social psychology has revealed that changing explicit biases alone is not sufficient to influence all gender biased decision and behavior, e.g., [15]–[17]. Studies of dual process models have shown that both conscious, controlled, reflective processes (such as explicit biases) and automatic associations that may be beyond conscious awareness (such as implicit biases) are essential forces shaping gender-biased decisions in workplaces [18].

However, SE literature on gender diversity has not yet paid enough attention to implicit gender biases. As the first step to addressing implicit gender biases, we need to develop empirically understandings of them in the software development profession. Hence, we ask the first main research question:

RQ_1 *Do implicit gender biases exist among software development professionals?*

If the implicit gender biases do exist among software development professionals, we want to examine if they significantly influence software development professionals' decision-making. Particularly, we focus on assessing job candidates and evaluating contributions. The literature has identified these two as the critical decision situations where gender-based discrimination often happens, e.g., [18]–[20].

RQ_2 *How implicit gender biases influence software development professionals' decision-making in assessing job candidates and evaluating contributions?*

To answer the above two research questions, we designed and conducted an empirical study with 142 professional software engineers from seven organizations. Each subject participated in a series of decision-making tasks. We also measured his or her implicit gender biases using three implicit association tests (IATs). The data analyses yield a rich set of findings. We found the evidence for the existence of strong implicit gender biases among both male and female software development professionals. These implicit gender biases shape their decisions on assessing job candidates and evaluating contributions. Our findings thus highlight the importance and urgency of addressing implicit gender biases in the software development profession. We discuss two potential approaches to cope with implicit gender biases at both individual and organizational levels. Specifically, the contributions of the study are four-fold:

- We conduct a well-designed empirical study. It does not only show the pervasiveness of the implicit gender bias among professional developers but also demonstrates that the implicit gender biases do impact people's decisions

[1]https://diversity.google/
[2]https://www-03.ibm.com/employment/inclusion/

978-1-7281-1762-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICSE-SEIS.2019.00009

on assessing job candidates and evaluating technical contributions.
- We introduce the concept of implicit gender bias in SE research. It will provide a new theoretical perspective and tool to SE researchers who are interested in gender inclusions in the software development profession.
- We develop two new domain-specific SE-related implicit attitude tests (IATs) for use in future SE research. As far as our current knowledge, it is the first attempt of developing SE-related IATs.
- We propose two potential measures to address the implicit gender bias at both individual and organizational levels.

The rest of the paper proceeds as follows. Section II introduces the background literature on implicit biases. Section III presents the study design. Section IV shows how we analyzed the data. Section V presents the results and findings. Section VI discusses the related issues. Section VII briefly reviews the related work. Section VIII concludes the paper.

II. Background

This section introduces the background of implicit gender bias. We start with the basic ideas and then explain how to measure it with the Gender-Career IAT as an example.

A. Implicit Bias

The concept of "implicit bias" was initially proposed in social cognition literature to understand this phenomenon [21]–[23]. As its name indicates, implicit bias refers to the attitudes or stereotypes that impact people's understanding, actions, and decisions in an unconscious manner. These biases, which encompass both favorable and unfavorable assessments, are activated involuntarily and without an individual's awareness or intentional control. Implicit bias and bias its corresponding measuring methods are often among the most significant progress in psychology [24]. Implicit gender biases have been proven to have strong impacts in many areas, for example, legal [25], politics [26], business & management [27], professional sports [28], health [29], and STEM [30]. It is reasonable to assume that implicit gender biases should play an important role in shaping people's decisions in the software development profession where gender stereotypic beliefs are quite strong [31].

B. Measuring Implicit Bias–Implicit Attitude Tests (IATs)

The IATs measure implicit biases in a simple yet compelling way. It asks participants to categorize information as quickly as possible and then calculates a participant's reaction time (in milliseconds) and the accuracy in completing the categorization task [22]. The rationale behind the IATs is that it should take less reaction time (i.e., response latency) make the same behavioral response (a key press) to concepts that are strongly associated than to concepts that are weakly associated in human cognitions [32]. Let us take the classic Gender-Career IAT deals with {*Male–Female*, *Career-Home*} [33] as an example. Following the conventions in Greenwald et al. [22], we name the {*Career-Home*} as target concept dimension, and {*Male-female*} as the attribute dimension. Obviously, if a person holds implicit gender biases, the concepts "*Career*" and "*Male*" tend to be more strongly associated than the concepts "*Career*" and "*Female*." Therefore, respondents should be able to identify and categorize items faster in a condition in which items representing "*Career*" and "*Male*" share the same response compared to a condition in which items representing "'*Career*" and "*Female*" share the same response. The situation is similar for the concept "Home."

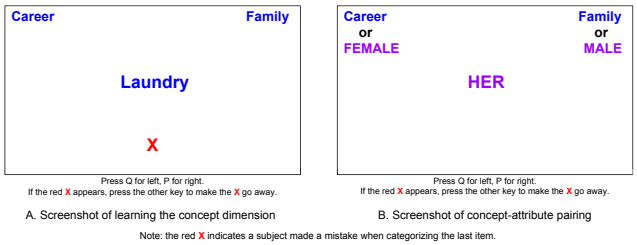

Fig. 1. The screenshot of the Gender-Career IAT's $Step_1$ and $Step_3$.

1) The Procedures of an IAT: A typical IAT consists of five steps [23], [32], [33]. Fig. 1 shows the screenshot of $Step_1$ and $Step_3$. The five steps are:

$Step_1$. Learning the target concept dimension–First, subjects sort items from two different concepts into their superordinate categories (e.g., word "salary" for *Career*, and word "garden" for *Home*). Subjects use two keys on a computer keyboard to indicate the categorization (see Fig. 1-A). In our study, we used "Q" and "P" because their positions on the keyboard are very convenient.

$Step_2$. *Learning the attribute dimension.* This step is similar to the $Step_1$. The attribute dimensions "*Male*" and "*Female*" also associate with a set of items (e.g.,"he" for *Male*, word "lady" for *Female*).

$Step_3$. *Concept-attribute pairing–round 1* This step combines two sorting tasks are combined such that, on alternating trials, respondents are identifying a word as *Career* or *Home* and then a word as *Male* or *Female*. Figure In this case, one key ("Q') is the correct response for two categories (*Career* and *Female*, see the top-left corner in Fig. 1-B) and the other key ("P") is the right response for the other two categories (*Home* and *Male*, see the top-left corner in Fig. 1-B). Subjects first perform a block of 20 trials to get familiar with these sorting rules (the "practice" block). After a brief pause, they continue for the second block of 40 trials (the "critical" block).

$Step_4$. Learning to switch the spatial location of the concepts. In the $Step_4$, stimulus items for the target concepts are sorted for 20 trials, but this time the key assignment is reversed. In our study, *Male* items would now require a "P" key response and *Female* items would require an "Q" key response.

$Step_5$. *Concept-attribute pairing–round 2*. In the $Step_5$, subjects sort items from both the attribute and target concept categories again. However, the response key assignments now require *Career* and *Male* items to be categorized with one key and *Home* and *Female* items to be categorized with the

other key, the opposite association from the $Step_3$. Subjects sort stimulus items with for 20 trials "practice" block and then again for 40 more trials "critical" block.

Note that the Gender-Career IAT only uses word stimulus, e.g., salary, garden, her, etc. Indeed, the stimulus could be in other forms such as pictures, for example, female faces for *Female*, or office desk for *career*.

2) Computing the IAT Effect: The IAT effect is calculated using latency (subjects' reaction time) data from $Step_3$ and $Step_5$'s critical tasks. The D-score is the most frequently used one. Greenwald et al. (2003) describe the algorithm for calculating the D-scores (Eq. 1).

$$D = \frac{\overline{L_3} - \overline{L_5}}{\sigma_{L_3 \cup L_5}} \quad (1)$$

It first calculates the difference in average response latency between the $Step_3$ and $Step_5$'s critical tasks and dividing by the standard deviation of all latencies for both sorting tasks. A D-score is bounded in [-2, 2]. The larger $|D|$ is, the more strong the implicit bias is. In general, the thresholds for "slight" (.15), "moderate" (.35) and "strong" (.65) were selected according to conventions for effect size [33].

III. Research Design

A. Subjects

We recruited subjects from the seven local software development organizations which had established collaboration with the authors' institution. When selecting participants, we kept the gender balance in the sample. In total, we recruited 160 subjects. However, we dropped 15 subjects' data for they had not satisfied the requirement of living in the United States for the last five years at the time of conducting this study[3]. Three subjects' data was also excluded because they subjects successfully suspected the real purpose of the study at the end of the decision-making tasks (after step 3 in Fig. 2).

In the remaining 142 subjects, 67 were male, and 75 were female. The population was also ethnically diverse. 65 (45.77%) identified themselves as Caucasian, 42 (29.58%) as Asian/Pacific Islands, 23 (16.20%) as Latino/Hispanic, 8 (5.63%) as African American, and 4 (2.82%) as multi-racial or Other. They held diverse positions in their organizations, for example, software architect, software engineer, QA engineer, project manager, front-end engineer, and so on. The mean age of the subjects was 33.72, and the average experience is 12.92 years. Most of the subjects (136 out of 142: 95.77%) held bachelor or above degrees. Two subjects were in the process of obtaining their degrees, and four held associate degrees or professional certificates.

Each subject received a $50 Amazon gift card as the compensation for participating the study[4]. Subjects did not need to commute to the authors' institution to participate in the study; instead, the first author and a student helper visited their organizations and conducted the study in their locations. We built a computerized study environment on a high-performance MacBook Pro laptop. All subjects finished the study using the same environment. Because the new MacBook Pro's keyboard is more prone to obstruction and the study required frequent keyboard hit, we provided an external Apple magic keyboard to improve subjects' experience during the study.

[3]This requirement is enforced to ensure subjects have enough exposure to social environment and culture in the United State

[4]The compensation was calculated using the average after-tax hour rate for software engineers in the region of the study. The hourly rate was based on the salary data from the Glassdoor.com.

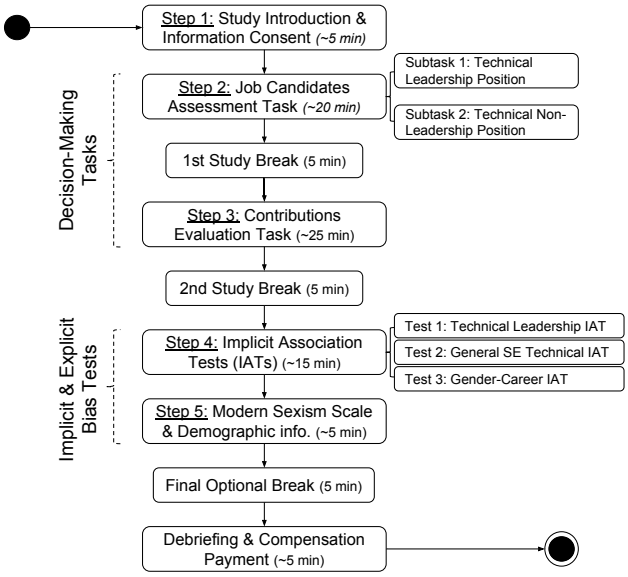

Fig. 2. Overview of the study process.

B. Study Procedures

The flowchart (Fig. 2) depicts the study procedures. The study consists of five major steps, two mandatory study breaks for subjects to have a rest, one optional study break, and the final debriefing session. During each study break, we provided light refreshments including water, soda, and snacks. The entire study process lasted about 90 minutes. Given the nature and the length of the study, we also allowed subjects to leave the room during the breaks.

Please note that the two decision tasks (step 2 & 3) happened before measuring gender biases (step 4 & 5). This arrangement is different from the order of our two research questions (RQ_1 and RQ_2). Doing so enables us to minimize the possibility for our subjects to suspect the intention of the study. Prior literature in gender bias has shown that people often deliberately make gender-neutral or lean-to-the-females decisions if they know the study is about gender inequality [25]. They do that because they want to show that they are not gender-biased. If we measure gender biases before the two decision-making tasks, it would be very likely that our subjects would assume the study is about gender issues and make decisions favoring the females. To control this factor, we put the two decision tasks before measuring gender biases. This allows us to collect people's undistorted decisions.

Before conducting the real study, we recruited five graduate students in the first authors' institution to perform a pilot study. We made some slight adjustments (e.g., short video tutorials on how to use the computerized study environment to perform each task) based on their feedback. The pilot study also demonstrated the interventions in the study is effective.

1) Step 1: Study Introduction & Information Consent: The first step focused on introducing the study procedures and collect subjects' consent. We also introduced the computerized study environment designed for the study. In order to keep the accuracy of the results, we did not disclose the real purpose of the study. Meanwhile, to avoid the ethical dilemma and the second-guessing effect [34], we did not use deception by telling a fake study purpose either. At the end of step 2, we asked each subject's opinion about the study purpose. Most of them except three subjects did not figure out the real purpose. Those three subjects' data was excluded though they still finished the entire study. We revealed the real purpose during the debriefing session at the end of the study.

2) Step 2: Job Candidates Assessment Task: The Job Candidates Assessment task has two subtasks. The first subtask is about evaluating candidates for a technical leadership position–Lead Software Architect, which is one of the most influential roles in software development. The second subtask is about evaluating candidates for a general technical non-leadership position–Full Stack Software Engineer.

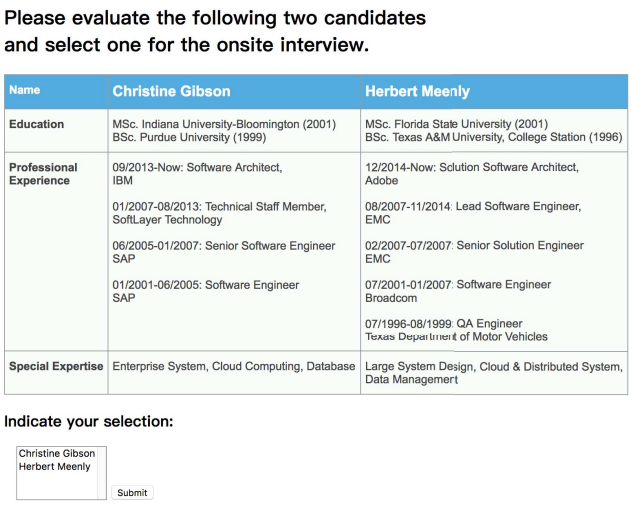

Fig. 3. The screenshot of decision-making task of selecting interview candidate for a lead software architect position. Each pair of candidates have almost equivalent qualifications regarding their education, experience, and expertise. Their names indicate their gender identities.

Subtask 1–Technical Leadership Position: At the beginning of the subtask, the study environment first presented a short video to introduce the position and its requirements. Then, each subject performed ten rounds of candidates evaluation. In each round, two candidates' profiles were shown side by side on the screen (Fig. 3). The subject was asked to select one for the next round onsite interview. In each round, the two candidates' qualifications are very similar. For example, in the screenshot (Fig. 3), both candidates' education backgrounds and work experiences are comparable. They all earned master degrees from large public research universities, served "software architect" roles in reputable software development organizations. The only significant differences are their genders (indicated by their names). *Christine Gibson* is a female name, while *Herbert Meenly* is a male name. All names are typical Caucasian names to avoid validity issues caused by the racial effect.

Three rounds had one female and one male candidate. The other seven rounds only had two same gender candidates. The seven same-gender rounds were not used in measuring subjects' decision tendency. They were used to prevent subjects from suspecting the real purpose of the study. The order of rounds and two candidates in each round (left/right) were randomized for the validity of the results.

Subtask 2–General Technical Position: The process of subtask 2 is similar to the process of subtask 1. The only differences are the fictitious job description and candidates' profiles. To keep the paper concise, we omit the detailed description of it.

3) Step 3: Contributions Evaluation Task: In the contributions evaluation task, the study environment first told subjects that they would review two software developers' attempts in fixing the same simple issue in each round. They are required to select the one that they were willing to accept. The fixes were simple; most were less than 5 lines of the code. The study used `Python` language. In each round, two software developers' fixes were displayed side by side with their names which made their genders explicit. The two fixes were semantically same. The differences could only be variable/method names (e.g., `get_auth_token()` vs. `get_authorization_token()`), or different implementation ways (e.g., different iteration conditions of a `for` loop).

Similar to the job candidates assessment tasks, we had three rounds of contributions evaluation with female and male developers. The rest seven rounds were with two developers of the same gender. Again, the seven same-gender rounds were used to avoid subjects figuring out the study purpose. Each subject's decisions in the three different gender rounds were used to measure his or her decision tendency.

4) Step 4: Gender Implicit Association Tests: The fourth step is to perform IATs to capture subjects' implicit gender biases. We performed three IATs. Two were domain-specific SE-related IATs developed by the research team, and the other is the standard Gender-Career IAT introduced in II-B. The word stimulus used in all three IATs are provided in the APPENDIX and free for use by any researchers.

To develop domain-specific SE-related IATs, we held a focus group meeting. Three software engineers, two SE researchers, and one psychologist participated in the meeting. For each of the two categories *Architect* and *SE Technical*, each software engineer provided a list of 30 words that he or she thought related to the category. Then, the two researchers sorted their words and identified the common words at least mentioned by two software engineers. From each category's common words, 7 were selected with the psychologist as the word stimulus for each IAT. The focus group meeting did

not identify the word stimulus for the category "Non-SE." We identified them using another procedure. We identified the most recent (as the date of 01/29/2018) 500 Non-SE jobs posted on Linkedin.com, extracted most common words in the job title, and compiled a 7-word list for the General SE Technical IAT.

All three IATs were implemented with the Open Source Implicit Association Test [35] developed by Ian Hussey at the Universiteit Gent, Belgium. We used our domain-specific word stimulus replaced the default ones, and slightly adjusted the user interfaces to improve the presentation of the information of concept categories.

5) Step 5: Modern Sexism Scale: In step 5, subjects also completed a Modern Sexism Scale (MSS) [36]. This scale is designed to measure explicit gender biases towards women in post cold-war western societies. The scale asks participants to respond to statements, such as "Over the past few years, the government and news media have been showing more concern about the treatment of women than is warranted by women's actual experiences." Participants respond to eight such statements on a numerical scale (strongly disagree - strongly agree: [1, 5]). Then, their responses are converted into a final score. Regarding this non-IAT measure, we can use it to examine whether there are correlations between IATs and explicit gender biases, as well as to fully estimate implicit gender biases' effects on decision-making in regression models. Besides, we can compare which measure (IATs or MSS) is better on predicting decision-making tendencies.

We also collected subjects' basic demographics information in this step. We asked subjects' education backgrounds, age, experience in professional software development, current position, political orientation, religiosity, sex-orientation, marriage status, residential environment, family size, gender identity, residential history in the US, and so on.

C. Measurements

1) Decision-making Tendencies: For all decision making tasks (Step 2 & 3: technical leadership, general SE, and contributions evaluation), we measure subjects' decision-making tendencies as follows.

Each decision task contained three rounds where a subject needed to make decisions between a man and a woman. For the three rounds, his or her choices could be one of the following four cases: (a)–3 women, (b)–2 women and 1 man, (c)–1 woman and 2 men, (d)–3 men. Since there is no significant differences between the two candidates' qualifications and the display orders were also randomized, we can assume their choices reflect their preferences on a specific gender. Obviously, the case (a) indicates strongly prefer female candidates or their work, while case (d) indicates strongly prefer male candidates or their work. Cases (b) and (c) are in the middle. Thus, we coded each subject's choices in each task into a four-level ordinal variable representing his or her gender-related decision-making tendency in the task. We mapped (a), (b), (c), (d) to numerical values 4, 3, 2, 1. The larger it is, the more likely that a subject tends to favor a female candidate or her work, and vice versa.

2) IATs and MSS: These constructs are measured by the standard measurements. For IATs, we recorded each's raw latency data in milliseconds. We then calculated the D-scores using the method introduced in II-B. We followed the standard process to score MSS responses.

IV. Data Analysis

We employed multiple statistical techniques to analyze the data. All statistical analyses are performed with `R 3.4.1` [37] and its associated packages for MAC OS High Sierra (version 10.13.1). We follow the ASA's principles to present and interpret the statistical significance [38].

A. RQ_1: D-Score and Welch's Two Sample t-test

Answering RQ_1 is straightforward. To test the existence of implicit gender biases, we used the D-score, which is the standard measure for an IAT. We also tested the subjects' reaction time difference when pairing two associations. In our study, we used Welch's Two Sample *t*-test which is more robust and reliable than student t-test when the two samples have different variances [39]. Besides, as a validity check, we also want to examine the relationship between IATs (implicit gender biases) and MSS (explicit gender bias). We use the correlation analysis to test it.

Another issue to explore is if both female and male subjects exhibit similar levels of implicit gender biases. To investigate it, we divided our sample into two samples: *male subjects*, and *female subjects*. For each IAT, we calculated D-scores for both gender groups.

B. RQ_2: Generalized Linear Modeling (GLM)

Answering RQ_2 requires demonstrating that the implicit gender biases strongly impact the people's decisions on evaluating job candidates and contributions. Simple correlation analysis may not be sufficient because significant correlations do not guarantee the link between implicit gender biases and decisions. We employed the Generalized Linear Modeling (GLM) technique to perform the analyses [40]. Obviously, the decision tendencies on evaluating job candidates and contributions are *dependent variables* (see Section III-C1 for the coding schema). IATs' results are *independent variables*. We used each subject's absolute latency differences (in millisecond) between two pairing tasks ($Step_3$ and $Step_5$, see section II-B) as its value. Given that establishing relationships between implicit gender biases and decisions requires excluding the effects of explicit gender biases, each subject's MSS result is treated as a *control variable*. The demographic characteristics collected in Section III-B5 are also *control variables*.

V. Results and Findings

In this section, we report the main findings of the study. We organize the results according to the corresponding research questions. By convention, the D-score's thresholds for "slight," "moderate," and "strong" implicit biases are 0.15, 0.35, and 0.65 respectively.

A. RQ_1: Gender Implicit Bias

1) IAT Test 1: Technical Leadership: The Technical Leadership IAT shows that software development professionals hold implicit gender biases related to technical leadership positions (in our study, software architect) in the software development profession. Subjects' reactions to stimulus displayed an association between *Architect* and *Male* (Mean=784.20) compared to *Architect* and *Female* (Mean=1055.13). The D-score is 1.12, indicating a **strong** effect of implicit gender biases. Welch's Two Sample *t*-test on two concept-pairings are: $t = 19.575$, $df = 242.74$, $p < .001$.

2) IAT Test 2: General SE Technical: The General SE Technical IAT shows that software development professionals hold implicit gender biases related to general technical positions in the software development profession. Subjects' reactions to stimulus displayed an association between *SE-Technical* and *Male* (Mean=815.49) compared to *SE-Technical* and *Female* (Mean=963.81). The D-score is 0.60, indicating a **medium** effect of implicit gender biases. and Welch Two Sample t-test on two concept-pairings are: $t = 5.29$, $df = 260.38$, $p < .001$.

3) IAT Test 3: Gender-Career: The Gender-Career IAT shows that software development professionals hold implicit gender biases connecting women with the home/family, as well as men with career/work. Subjects' reactions to stimulus displayed an association between *Career* and *Male* (Mean=759.16) compared to *Career* and *Female* (Mean=1078.43). The D-score is 0.73, indicating a **strong** effect of implicit gender biases. Welch's Two Sample t-test on two concept-pairings are: $t = 13.571$, $df = 267.59$, $p < .001$.

4) IATs vs. MSS: We further tested if the results of any of the three IATs would be related to the results of the MSS for explicit gender biases. The analyses show that none of the IATs were correlated with the MSS at the .05 significant level, which indicates that IATs and MSS do measure different constructs. The differences between the results of the IATs and the MSS suggest that investigating both implicit gender biases and explicit gender attitudes is necessary and important.

5) Female vs. Male: Tab. I lists the D-scores of each gender group in the three IATs. Both gender groups have implicit gender biases though there some slight differences regarding the degree of them. We want to highlight an interesting but sad finding. Let us have a look at the the row of "Technical Leadership." The female subjects' D-score of the Technical Leadership IAT is higher than the males' (1.27 vs. 1.09). Female subjects even hold stronger implicit gender bias towards female leaders than male subjects do. It look like that women even more polluted by the social beliefs that men should be leaders in the software development profession.

TABLE I
D-SCORES OF EACH GENDER GROUP IN IATS.

IATs	Female	Male	Both
Technical Leadership	1.27	1.09	1.12
General SE Technical	0.43	0.78	0.60
Gender-Career	0.67	0.84	0.73

6) Answers to RQ_1: Based on the above results, we can answer RQ_1 as follows:

> *Implicit gender biases are widely existing among professional software developers. For all three IATs, the effects of associating men with Technical Leadership (Architect), General SE Technical (SE Technical), and Career are significant for both male and female subjects.*

B. RQ_2: Implicit Bias and Decision Making

We present the GLM results in the Tab. II. Models T_0, G_0, and C_0 are the models that only have demographic control variables and the MSS (explicit gender bias score). Models T_I, G_I, and C_I are the models incorporating three IATs' results. It is easy to find that models X_I consistently have lower *AIC (Akaike Information Criterion)* and higher *MAE (Mean Absolute Error)*; see the Δ_{AIC} and Δ_{MAE} in Tab. II for details. Therefore, adding the results of the three IATs does produce better models for predicting software development professionals' decision tendencies on job candidates assessment and contributions evaluation. Another interesting result is that the impact of explicit gender biases is limited. In all models, MSS are either not significant or only marginally significant ($0.05 \leq p < 0.1$), indicating people now can (at least partially) resist their explicit gender biases.

1) Task 1: Job Candidates Assessment: Columns 2 to column 5 of Tab. II show the results of the job candidates assessment. Column 2 and 3 are for the subtask 1, and column 4 and 5 are for subtask 2. In Models T_0 and G_0, MSS are marginally significant. After adding three implicit gender biases, MSS are no longer significant in T_1 while all three IATs' results become significant in Models T_1 and G_1. The increases on implicit gender biases reduce the possibility of making decisions favoring the females (negative βs). The only difference is: for subtask 1–Technical Leadership position (Lead Software Architect), results of Technical Leadership IAT has a more significant effect ($\beta = -0.015$); for subtask 2–General SE position (Full Stack Software Engineer), results of General SE Technical IAT has a more significant effect ($\beta = -0.024$). It also provides some indirect evidence for the validity of the two domain-specific IATs.

2) Task 2: Contributions Evaluation: Columns 6 and 7 of Tab. II show the results of the contributions evaluation. The results are similar to the Job Candidates Assessment. However, the impact of IATs seems smaller in this task through the results of General SE Technical IAT is still significant. The gain of *MAE* is also smaller. There are may be multiple reasons for this. One possible explaination is that the job candidates assessment task asked subjects to direct compare people, while the contributions evaluation task did not require direct people evaluations. Hence, they may be able to make more gender-neutral decisions. However, future studies will help us to understand this issue better.

3) Answers to RQ_2: Based on the above results, we can answer RQ_1 as follows:

TABLE II
GLM ANALYSIS RESULTS.

	Job Candidates Assessment				**Contributions Evaluation**	
	Subtask 1-Technical Leadership Position		**Subtask 2-General SE Technical Position**			
Var.	Model T_0 (β)	Model T_I (β)	Model G_0 (β)	Model G_I (β)	Model C_0 (β)	Model C_I (β)
Demographic Controls†	$-.-^{***}$	$-.-^{***}$	$-.-^{***}$	$-.-^{***}$	$-.-^{***}$	$-.-^{***}$
Explicit Gender Biases						
MSS	$-0.241*$	-0.093	$-0.165*$	$-0.139*$	-0.087	-0.046
Implicit Gender Biases						
Technical Leadership		-0.015^{***}		$-0.009*$		$-0.007*$
General SE Technical		$-0.002**$		-0.024^{***}		$-0.013**$
Gender-Career		$-0.008*$		$-0.010**$		$-0.021*$
AIC	1204.58	829.72	1449.63	1282.45	1215.31	1096.29
(Δ_{AIC})	$\Delta_{AIC} = 374.86$		$\Delta_{AIC} = 167.18$		$\Delta_{AIC} = 119.02$	
MAE	0.69	0.26	0.54	0.29	0.33	0.21
(Δ_{MAE})	$\Delta_{MAE} = 0.43$		$\Delta_{MAE} = 0.25$		$\Delta_{MAE} = 0.12$	
No. of Observations	142					

†: Given we have over a dozen of demographic control variables (see section III-B5), we use "demographic controls" to represent them without showing details; the $*$ signs indicates the significance levels of models with only demographics control variables and the MSS.
$*$: $p < 0.10$, $**$: $p < 0.05$, $***$: $p < 0.01$.

Implicit gender biases have significant impacts on professional software developers' decision-making on job candidates assessment and contributions evaluation. Specifically, while people can resist the impact of explicit gender biases, their decisions are still largely influenced by implicit gender biases.

VI. DISCUSSION

A. Addressing Implicit Gender Biases

The findings of our empirical study paint a not-so-optimistic picture of implicit gender bias in the software development profession. What concern us are that the study participants consistently held implicit gender biases, and these biases influenced their decisions. Thus, the implicit gender biases are very likely to put female software engineers in disadvantageous situations. Thus, remedies are highly necessary to counter the unfavorable effects of implicit gender biases in software development organizations.

To reduce the implicit gender biases, we propose two potential methods at both the individual and organizational levels. First, designing a series of carefully crafted and empirically tested bias reduction training courses in both educational institutions and software development organizations. The trainings may include a component that let trainees better understand themselves through direct confronting them with their own implicit biases [41]. Doing so has the potential to reduce implicit gender biases of both current professional and SE students (future professionals). The continuous trainings hence maximizes the effects. However, most bias reduction strategies are only temporary measures. One cannot expect a simple training, or even a few sparse courses, to come close to permanently reversing the harmful effects of implicit gender bias. Indeed, it is likely that the benefits gained even by a continuous training program would be limited. Thus, the training must be regular and continue throughout careers, at least for those who are high on implicit bias. The second method is at the organizational level. Software development organizations may proactively hire women in counter-stereotypic roles, and encourage their female employees to take technical leadership roles. Having more female technical leaders does not only help reduce implicit biases but also challenge and transform the male-dominated organizational/societal cultures [42].

B. Threats to Validity

As any empirical studies, our study is not free of threats to validity. We briefly discuss them from three perspectives.

First, from the perspective of construct validity, there is no significant threat. The central construct of the study is the implicit gender biases. We used the Implicit Association Tests (IATs) to measure them from multiple perspectives. The validity of IATs has been confirmed by thousands of studies in many disciplines, for example, management, psychology, sociology, etc., and also have been finished by millions of participants. Our domain-specific SE IATs were developed through strict procedures with the involvement of a professional psychologist. Another construct, explicit gender bias, was measured using Modern Sexism Scales which is also a widely adopted, reliable measure [36]. Using standard psychometric models enhances our confidence in construct validity.

Second, from the perspective of internal validity, we took multiple measures to remove most of them when performing the study. For example, we added the same gender pairs in the hiring decision task, to effectively prevent subjects to figure out the real intention of the study. Decision tasks were arranged before the IATs and MSS to avoid the learning effects. We also excluded data from the subjects who managed to suspect the true study purpose at the end of the decision-making tasks. For more details, please refer to the Section III-B. However, we must admit that we cannot control all factors. For example, in the hiring decision-making task, a

subject may find he or she graduated from the same university or worked for the same company with a fictitious candidate. The alumni relationships may make a subject favor that candidate. However, no social experiment can perfectly control every factor [43]. At least, the randomness we introduced to the study should help to alleviate such threats.

Third, from the perspective of external validity, our study was conducted with software development organizations located in the region of Western New York which is assumed to be a moderate liberal area regarding culture and politics[5]. Hence, we may be able to reach a proposition that the results at least valid for the regions sharing the same characteristics. In more liberal regions, e.g., New York City or Bay area, the problem of implicit gender bias may be less severe. In more conservative regions, e.g., Bible belt in the Southern United States, implicit gender bias, or even explicit gender bias may be more significant issues. Further replication studies with software organizations will paint a more comprehensive picture of implicit gender bias around the United States. Besides, replications in other countries will also bring an additional understanding of implicit gender bias, which is indeed a global issue.

VII. Related Work

Researchers have done a large amount of work on identifying gender gaps in computing and promoting women's participation. We briefly reviewed some work in both research streams. The study described in this paper is different from the reviewed work in three aspects. First, our study focuses on capturing professional software engineers' unconscious implicit gender biases. Doing so enables other researchers and us to understand people's gender attitude from a more comprehensive perspective. As far as our current knowledge, our study is the first work that investigates gender equality through the lens of implicit bias in the SE domain. Second, our study focuses on the social cognitions which are root causes of the gender gaps in SE.

A. Gender Gaps in Computing

Recent research has documented that women are underrepresented in the software development profession and suffer from various types of prejudice and bias [44]. Beyond the statistics listed in the introduction, formal research also confirms the situation that women are underrepresented in the software development industry, and in almost all major economies. Ironically, information technology, which was viewed as an opportunity for women in the early 2000s, reproduces and strengthen gender inequalities seen in the broader fabric of society [45], [46].In addition to the offline setting, researchers have investigated gender and racial inequalities in emerging online communities and labor markets. Their results unanimously suggest the fact that the gender gap is just as prevalent and relevant online as offline, for example, [47]–[51]. These studies also focus on different sources of bias. For instances, Terrell et al. identify individual-level explicit biases based on information available on GitHub users' profiles [3]; while Stephens [52] uses Google Map as an example to show favoring men's tastes at the community level. Besides, some studies such as [53] also highlights subtle yet systematic types of biases resulting from using algorithms and its potential legal risks. There is also some research focusing on the labor market protections in the online Gig economics [49].

Specifically, Stack Overflow has been a well-studied community for gender-related issues. Vasilescu et al. show that women are underrepresented in this community [54]. Several significant barriers to women's participation are identified through qualitative interviews. Such barriers include the lack of awareness of some site features, the intimidating community size, fear of receiving negative/unfriendly comments and lacking confidences on their qualifications. Ford et al. [55] show that a female user's participation would improve if another woman appears in the same question thread. There are also a few other studies such as [56], [57] depict similar pictures.

B. Promoting Women's Participation

To help remedy this situation, researchers have done a great deal of work to promote women's participation in software development. For example, Gorriz and Medina [58] suggested that software games can promote women's engagement in computing. Gweon et al. [59] explores design concepts and principles to engage girls in learning programming through different media and interaction techniques. Studies such as [60] attempt to identify the social and psychological factors deterring women's participation. Retention of female developers is also often discussed as an issue [61].

Researchers have developed engineering approaches to help improve the inclusiveness of women in computing. Researchers has developed a family of Gender HCI techniques to make the end-user development more gender inclusive, e.g., [7]. GenderMag approach was proposed and applied it to software product teams in Microsoft [62]. Fairness testing, a method of testing software for discrimination, may also help with challenges of inclusiveness although it was designed for detecting multiple types of biases [63].

There is empirical evidence demonstrating that increasing gender diversity in software development teams also leads to practical values. Vasilescu et al. [4] shows that gender diversity has a positive influence on a team's productivity. Aue et al. [64] suggests similar results that social diversity does contribute to project growth. While these studies prove the benefits of increasing gender diversity, a reluctant fact is that females' participation is still quite peripheral [65]. Moreover, a recent study confirms that over 80% of the barrier types for newcomer include attributes that are biased against women [9]. Since women often suffer the competence-confidence gap [66], creating a path and lowering barriers for female developers' self-guided personal development is critical for them to achieve their technical distinctions and for the community to benefit from gender diversity.

[5]https://en.wikipedia.org/wiki/Politics_of_Upstate_New_York

VIII. Conclusion and Future Work

With the progress of the mordern society, some people tend to believe that the females and males are treated almost equally now. However, the real situations are not that optimistic. While explicit gender biases are decreasing, the biases towards the females become more subtle and unconscious. The implicit gender biases still significantly impact people's decision and behavior. The software development profession, where gender diversity is a long-lasting problem, is not free of such implicit gender biases.

In this paper, we describe a study that empirically examines the implicit gender biases in the software development profession. With data collected from 142 professional software engineers in seven organizations, our study reveals rich evidence and understandings of implicit gender biases and their effects on people's decision-making. Our findings show a concerning yet hopeful picture: first, implicit gender biases widely exist among both female and male professional software developers and have strong impacts on people's decisions on job candidates assessment and contributions evaluation; however, a good sign is people now can resist the impacts of explicit gender biases in decision-making.

Although our research provided detailed information about implicit gender biases in the software development profession, further empirical inquiries should continue to investigate implicit gender bias to gain more complete understandings of implicit gender biases and their effects in critical decisions in software development. In addition to job candidates and contributions evaluation, there are many other situations that implicit gender biases may play an important role, for example, assigning challenging technical work to team members. Future empirical work in this direction is promising. Identifying practical approaches to address implicit gender biases in software development is also critical. Indeed, successfully dealing with implicit biases also signals significant culture and society progresses.

Appendix A
Word Stimulus for The Three IATs

Tab. III lists the word stimulus used in the three IATs. We directly used the Gender-Career IAT in [33]. The other two were domain-specific SE-related IATs.

TABLE III
The word stimulus for the three IATs.

IAT	Concept Category	Word Stimulus
Gender-Career	*Career*	office, manager, salary, job, briefcase, profession, employees
	Home	garden, kitchen, marriage, laundry, home, children, relatives
	*Male**	MAN, HE, MEN, HIM, BOY, HIS, GENT
	*Female**	WOMAN, SHE, WOMEN, HER, GIRL, HERS, LADY
Technical Leadership**	*Architect*	architectural style, design, delivery management, enterprise solutions, RESTful API Design, large scale internet application, microservices
	*SE Technical****	debug, testing, DevOps, compiler, IDE, version control, commit code
General SE Technical**	*SE Technical****	debug, testing, DevOps, compiler, IDE, version control, commit code
	Non-SE	marketing, designer, assistant, customer service, accountant, data entry, human resource

*. All word stimulus for *Male* and *Female* are capitalized in all IATs.
**. We reused word stimulus for *Male* and *Female* from the Gender-Career IAT in the Technical Leadership and General SE Technical IATs.
***. The word stimulus of *SE Technical* reused in the Technical Leadership and General SE Technical IATs.

References

[1] M. Ortu, G. Destefanis, S. Counsell, S. Swift, M. Marchesi, and R. Tonelli, "How diverse is your team? investigating gender and nationality diversity in github teams," PeerJ Preprints, Tech. Rep., 2016.

[2] A. Panteli, J. Stack, M. Atkinson, and H. Ramsay, "The status of women in the uk it industry: an empirical study," *European Journal of Information Systems*, vol. 8, no. 3, pp. 170–182, 1999.

[3] J. Terrell, A. Kofink, J. Middleton, C. Rainear, E. Murphy-Hill, C. Parnin, and J. Stallings, "Gender differences and bias in open source: Pull request acceptance of women versus men," *PeerJ Computer Science*, vol. 3, p. e111, 2017.

[4] B. Vasilescu, D. Posnett, B. Ray, M. G. van den Brand, A. Serebrenik, P. Devanbu, and V. Filkov, "Gender and tenure diversity in github teams," in *Proc. CHI '15*. ACM, 2015, pp. 3789–3798.

[5] B. Vasilescu, V. Filkov, and A. Serebrenik, "Perceptions of diversity on github: A user survey," in *Proc. CHASE'15*, 2015, pp. 50–56.

[6] datausa.io, "Demographics of software developers (2015)," https://datausa.io/profile/soc/15113X/, 2015, [Online; accessed 09/26/2017].

[7] L. Beckwith and M. Burnett, "Gender: An important factor in end-user programming environments?" in *Proc. VL/HCC'04*. IEEE, 2004, pp. 107–114.

[8] J. E. Fountain, "Constructing the information society: women, information technology, and design," *Technology in Society*, vol. 22, no. 1, pp. 45–62, 2000.

[9] C. Mendez, H. S. Padala, C. H. Zoe Steine-Hanson, A. Horvath, C. Hill, L. Simpson, N. Patil, A. Sarma, and M. Burnett, "Open source barriers to entry, revisited: A tools perspective," in *Proc. ICSE '18*, 2018, pp. Accepted, to appear.

[10] G. T. Richard, Y. B. Kafai, B. Adleberg, and O. Telhan, "Stitchfest: Diversifying a college hackathon to broaden participation and perceptions in computing," in *Proc. SIGCSE'15*, 2015, pp. 114–119.

[11] A. Roan and G. Whitehouse, "Women, information technology and 'waves of optimism': Australian evidence on 'mixed-skill' jobs," *New Technology, Work and Employment*, vol. 22, no. 1, pp. 21–33, 2007.

[12] A. Powell, B. Bagilhole, and A. Dainty, "How women engineers do and undo gender: Consequences for gender equality," *Gender, Work & Organization*, vol. 16, no. 4, pp. 411–428, 2009.

[13] S. Rogerson, J. Weckert, and C. Simpson, "An ethical review of information systems development–the australian computer society's code of ethics and ssadm," *Information technology & people*, vol. 13, no. 2, pp. 121–136, 2000.

[14] P. Tourani, B. Adams, and A. Serebrenik, "Code of conduct in open source projects," in *Proc. SANER*. IEEE, 2017, pp. 24–33.

[15] M. Bertrand, D. Chugh, and S. Mullainathan, "Implicit discrimination," *American Economic Review*, vol. 95, no. 2, pp. 94–98, 2005.

[16] C. L. Ridgeway and S. J. Correll, "Unpacking the gender system: A theoretical perspective on gender beliefs and social relations," *Gender & society*, vol. 18, no. 4, pp. 510–531, 2004.

[17] J. C. Ziegert and P. J. Hanges, "Employment discrimination: the role of implicit attitudes, motivation, and a climate for racial bias." *Journal of Applied Psychology*, vol. 90, no. 3, pp. 553–562, 2005.

[18] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman, "Science faculty's subtle gender biases favor male students," *Proceedings of the National Academy of Sciences*, vol. 109, no. 41, pp. 16 474–16 479, 2012.

[19] C. Glass and K. L. Minnotte, "Recruiting and hiring women in stem fields." *Journal of diversity in Higher Education*, vol. 3, no. 4, pp. 218–229, 2010.

[20] M.-T. Wang and J. Degol, "Motivational pathways to stem career choices: Using expectancy–value perspective to understand individual and gender differences in stem fields," *Developmental Review*, vol. 33, no. 4, pp. 304–340, 2013.

[21] A. G. Greenwald and M. R. Banaji, "Implicit social cognition: attitudes, self-esteem, and stereotypes." *Psychological Review*, vol. 102, no. 1, p. 4, 1995.

[22] A. G. Greenwald, D. E. McGhee, and J. L. Schwartz, "Measuring individual differences in implicit cognition: the implicit association test." *Journal of Personality and Social Psychology*, vol. 74, no. 6, pp. 1464–1480, 1998.

[23] A. G. Greenwald, B. A. Nosek, and M. R. Banaji, "Understanding and using the implicit association test: I. an improved scoring algorithm." *Journal of Personality and Social Ssychology*, vol. 85, no. 2, pp. 197–215, 2003.

[24] I. Ajzen, "Nature and operation of attitudes," *Annual Review of Psychology*, vol. 52, no. 1, pp. 27–58, 2001.

[25] J. D. Levinson and D. Young, "Implicit gender bias in the legal profession: An empirical study," *Duke J. Gender Law & Policy*, vol. 18, p. 1, 2010.

[26] L. Huddy and N. Terkildsen, "Gender stereotypes and the perception of male and female candidates," *American Journal of Political Science*, pp. 119–147, 1993.

[27] C. L. Hoyt and J. L. Burnette, "Gender bias in leader evaluations: Merging implicit theories and role congruity perspectives," *Personality and Social Psychology Bulletin*, vol. 39, no. 10, pp. 1306–1319, 2013.

[28] J. S. Fink, "Female athletes, women's sport, and the sport media commercial complex: Have we really "come a long way, baby"?" *Sport Management Review*, vol. 18, no. 3, pp. 331–342, 2015.

[29] B. A. Teachman and K. D. Brownell, "Implicit anti-fat bias among health professionals: is anyone immune?" *International Journal of Obesity*, vol. 25, no. 10, p. 1525, 2001.

[30] S. Cheryan, J. O. Siy, M. Vichayapai, B. J. Drury, and S. Kim, "Do female and male role models who embody stem stereotypes hinder women's anticipated success in stem?" *Social Psychological and Personality Science*, vol. 2, no. 6, pp. 656–664, 2011.

[31] A. Smeding, "Women in science, technology, engineering, and mathematics (stem): An investigation of their implicit gender stereotypes and stereotypes' connectedness to math performance," *Sex Roles*, vol. 67, no. 11-12, pp. 617–629, 2012.

[32] B. A. Nosek, A. G. Greenwald, and M. R. Banaji, "Understanding and using the implicit association test: Ii. method variables and construct validity," *Personality and Social Psychology Bulletin*, vol. 31, no. 2, pp. 166–180, 2005.

[33] B. A. Nosek, M. R. Banaji, and A. G. Greenwald, "Harvesting implicit group attitudes and beliefs from a demonstration web site." *Group Dynamics: Theory, Research, and Practice*, vol. 6, no. 1, p. 101, 2002.

[34] R. Hertwig and A. Ortmann, "Deception in experiments: Revisiting the arguments in its defense," *Ethics & Behavior*, vol. 18, no. 1, pp. 59–92, 2008.

[35] I. Hussey, "Open source implicit association test," May 2018. [Online]. Available: osf.io/jvq4q

[36] J. K. Swim, K. J. Aikin, W. S. Hall, and B. A. Hunter, "Sexism and racism: Old-fashioned and modern prejudices." *Journal of Personality and Social Psychology*, vol. 68, no. 2, pp. 199–214, 1995.

[37] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org

[38] R. L. Wasserstein and N. A. Lazar, "The asa's statement on p-values: Context, process, and purpose," *The American Statistician*, vol. 70, no. 2, pp. 129–133, 2016.

[39] B. Derrick, D. Toher, and P. White, "Why welch's test is type i error robust," *The Quantitative Methods in Psychology*, vol. 12, no. 1, pp. 30–38, 2016.

[40] P. McCullagh and J. A. Nelder, *Generalized linear models*. CRC press, 1989, vol. 37.

[41] A. M. Czopp, M. J. Monteith, and A. Y. Mark, "Standing up for a change: Reducing bias through interpersonal confrontation." *Journal of Personality and Social Psychology*, vol. 90, no. 5, p. 784, 2006.

[42] C. Staats, K. Capatosto, R. A. Wright, and D. Contractor, *State of the science: Implicit bias review 2015*. Kirwan Institute for the Study of Race and Ethnicity, The Ohio State University, 2015, vol. 3.

[43] R. E. Kirk, "Experimental design," *The Blackwell Encyclopedia of Sociology*, 2007.

[44] M. Meyer, A. Cimpian, and S.-J. Leslie, "Women are underrepresented in fields where success is believed to require brilliance," *Frontiers in Psychology*, vol. 6, 2015.

[45] M. Foschi, L. Lai, and K. Sigerson, "Gender and double standards in the assessment of job applicants," *Social Psychology Quarterly*, pp. 326–339, 1994.

[46] E. Ruiz Ben, "Defining expertise in software development while doing gender," *Gender, Work & Organization*, vol. 14, no. 4, pp. 312–332, 2007.

[47] A. Hannák, C. Wagner, D. Garcia, A. Mislove, M. Strohmaier, and C. Wilson, "Bias in online freelance marketplaces: Evidence from taskrabbit and fiverr," in *Proc. CSCW'17*. ACM, 2017, pp. 1914–1933.

[48] E. Graells-Garrido, M. Lalmas, and F. Menczer, "First women, second sex: Gender bias in wikipedia," in *Proc. HT'15*, 2015, pp. 165–174.

[49] S. Jia, T. Lansdall-Welfare, S. Sudhahar, C. Carter, and N. Cristianini, "Women are seen more than heard in online newspapers," *PloS One*, vol. 11, no. 2, p. e0148434, 2016.

[50] J. Wachs, A. Hannák, A. Vörös, and B. Daróczy, "Why do men get more attention? exploring factors behind success in an online design community," *arXiv preprint arXiv:1705.02972*, 2017.

[51] C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier, "It's a man's wikipedia? assessing gender inequality in an online encyclopedia." in *Proc. ICWSM'15*, 2015, pp. 454–463.

[52] M. Stephens, "Gender and the geoweb: divisions in the production of user-generated cartographic information," *GeoJournal*, vol. 78, no. 6, pp. 981–996, 2013.

[53] J. A. Kroll, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, "Accountable algorithms," *University of Pennsylvania Law Review*, vol. 165, p. 633, 2016.

[54] B. Vasilescu, A. Capiluppi, and A. Serebrenik, "Gender, representation and online participation: A quantitative study of stackoverflow," in *Proc. SocialInfo.'12*. OUP, Dec 2012, pp. 332–338.

[55] D. Ford, A. Harkins, and C. Parnin, "Someone like me: How does peer parity influence participation of women on stack overflow?" in *Proc. VL/HCC'17*, Oct 2017, pp. 239–243.

[56] B. Lin and A. Serebrenik, "Recognizing gender of stack overflow users," in *Proc. MSR'16*. ACM, 2016, pp. 425–429.

[57] S. Morgan, "How are programming questions from women received on stack overflow? a case study of peer parity," in *SPLASH Companion'17*. ACM, 2017, pp. 39–41.

[58] C. M. Gorriz and C. Medina, "Engaging girls with computers through software games," *Commun. ACM*, vol. 43, no. 1, pp. 42–49, Jan. 2000.

[59] G. Gweon, J. Ngai, and J. Rangos, "Exposing middle school girls to programming via creative tools," in *Proc. INTERACT'05*. Berlin, Heidelberg: Springer-Verlag, 2005, pp. 431–442.

[60] S. Cheryan, V. C. Plaut, P. G. Davies, and C. M. Steele, "Ambient belonging: how stereotypical cues impact gender participation in computer science." *Journal of Personality and Social Psychology*, vol. 97, no. 6, p. 1045, 2009.

[61] J. M. Cohoon, "Toward improving female retention in the computer science major," *Commun. ACM*, vol. 44, no. 5, pp. 108–114, May 2001.

[62] M. Burnett, S. Stumpf, J. Macbeth, S. Makri, L. Beckwith, I. Kwan, A. Peters, and W. Jernigan, "Gendermag: A method for evaluating software's gender inclusiveness," *Interacting with Computers*, vol. 28, no. 6, pp. 760–787, 2016.

[63] S. Galhotra, Y. Brun, and A. Meliou, "Fairness testing: Testing software for discrimination," in *Proc. ESEC/FSE'17*. ACM, 2017, pp. 498–510.

[64] J. Aué, M. Haisma, K. F. Tómasdóttir, and A. Bacchelli, "Social diversity and growth levels of open source software projects on github," in *Proc. ESEM'16*. ACM, 2016, pp. 41:1–41:6.

[65] C. Fiesler, S. Morrison, R. B. Shapiro, and A. S. Bruckman, "Growing their own: Legitimate peripheral participation for computational learning in an online fandom community," in *Proc. CSCW '17*. ACM, 2017, pp. 1375–1386.

[66] Z. Wang, Y. Wang, and D. Redmiles, "Competence-confidence gap: A threat to female developers' contribution on github," in *Proc. ICSE '18*, 2018, pp. Accepted, to appear.