Biases and Differences in Code Review Using Medical Imaging and Eye-Tracking: Genders, Humans, and Machines

Yu Huang\textsuperscript{1}, Kevin Leach\textsuperscript{1}, Zohreh Sharafi\textsuperscript{1}, Nicolas McKay\textsuperscript{1}, Tyler Santander\textsuperscript{2}, Westley Weimer\textsuperscript{1}

\textsuperscript{1}University of Michigan, Ann Arbor
\textsuperscript{2}University of California, Santa Barbara
Thank You to the Collaborators!

Yu Huang

Dr. Kevin Leach

Dr. Zohreh Sharafi

Nicholas McKay

Dr. Tyler Santander

Dr. Westley Weimer

Yu Huang @ FSE 2020
On the Academic Job Market!

Yu Huang
SE & Human Factors
yhhy@umich.edu

Dr. Kevin Leach
SE & Security
kjleach@umich.edu

Dr. Zohre Sharafi
Empirical SE & Human Aspects
zohrehsh@umich.edu
Motivation

- Code review is *critical* for software development
- *Systematic* inspection, analysis, evaluation, and revision of code.
Motivation

- Code review is *critical* for software development
  - *Systematic* inspection, analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be 60%-65%
Motivation

- Code review is critical for software development
  - Systematic inspection, analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be 60%-65%
Motivation

- Code review is **critical** for software development
  - **Systematic** inspection, analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be **60%-65%**
Motivation

- Code review is *critical* for software development
  - *Systematic* inspection, analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be 60%-65%

Commit message

Delete the equal mark in case the array is like \{x,x,x...\(n\),y,y,y...\(n+1\}\}

Code changes
Motivation

- Code review is *critical* for software development
  - *Systematic* inspection, analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be 60%-65%
Motivation

- Code review is critical for software development
  - Systematic inspection, analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be 60%-65%
Motivation

- Code review is **critical** for software development
  - **Systematic** inspection, analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be **60%-65%**
Motivation

- Code review is **critical** for software development
  - **Systematic** inspection, analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be **60%-65%**
Motivation

- Code review is *critical* for software development
  - *Systematic* inspection, analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be 60%-65%
Motivation

- Code review is critical for software development
  - Systematic inspection, analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be 60%-65%
Motivation

- Code review is **critical** for software development
  - *Systematic* inspection, analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be **60%-65%**
High-level Question

- Is there bias on gender and identities in code review? How do we characterize the bias?
High-level Question

- Is there bias on gender and identities in code review? How do we characterize the bias?
  - Systematically
  - Objectively
  - Rigorously
High-level Question

- Is there bias on gender and identities in code review? How do we characterize the bias?
  - Systematically
  - Objectively
  - Rigorously

Behavioral Differences
High-level Question

- Is there bias on gender and identities in code review? How do we characterize the bias?
  - Systematically
  - Objectively
  - Rigorously

Behavioral Differences

Visual Differences
High-level Question

- Is there bias on gender and identities in code review? How do we characterize the bias?
  - Systematically
  - Objectively
  - Rigorously

Behavioral Differences | Visual Differences | Neurological Differences
Outline

● Motivation
● High-level question
● Experimental design
● Results
● Conclusions
Experimental Design: Code Review Tasks

Code Reviews
Experimental Design: Code Review Tasks

- Behavioral Differences
- Visual Differences
- Neurological Differences
- Decision, Response Time
Experimental Design: Code Review Tasks

Code Reviews

Behavioral Differences → Decision, Response Time

Visual Differences → Eye Tracker

Neurological Differences
Experimental Design: Code Review Tasks

- Behavioral Differences
- Visual Differences
- Neurological Differences
- Decision, Response Time
- Eye Tracker

Yu Huang @ FSE 2020
Experimental Design: Code Review Tasks

- Behavioral Differences
- Visual Differences
- Neurological Differences

- Code Reviews
- Decision, Response Time
- Eye Tracker
- fMRI

Yu Huang @ FSE 2020
Experimental Design: Code Review Tasks

- Behavioral Differences
- Visual Differences
- Neurological Differences
- Code Reviews
- Decision, Response Time
- Eye Tracker
- fMRI
Experimental Design: Code Review Tasks

Code Reviews

Behavioral Differences → Decision, Response Time

Visual Differences → Eye Tracker

Neurological Differences → fMRI
Experimental Design: Code Review Tasks

- How to control the variables of authors except for genders?
  - Race
  - Age
  - Attractiveness
  - Facial expressions
Experimental Design: Code Review Tasks

● How to control the variables of authors except for genders?
  ○ Race
  ○ Age
  ○ Attractiveness
  ○ Facial expressions

● How to fit everything with the constraints of the experimental environment?
  ○ Limited time
  ○ Requirements for different measures
Experimental Design: Code Review Tasks

- **How to control the variables of authors *except for genders***?
  - Race
  - Age
  - Attractiveness
  - Facial expressions

- **How to fit everything with *the constraints* of the experimental environment**?
  - Limited time
  - Requirements for different measures

- **How to control code quality***?
Experimental Design: Code Review Tasks

- 60 C/C++ pull requests from GitHub
  - 20 adopted from a previous study
  - 40 from the top 60 starred C/C++ projects
Experimental Design: Code Review Tasks

- 60 C/C++ pull requests from GitHub
  - 20 adopted from a previous study
  - 40 from the top 60 starred C/C++ projects
- Author images: Relabel the author information
  - Human: man, woman
    - Chicago Face Database (CFD)
      - Age, race, attractiveness, facial expression
Experimental Design: Code Review Tasks

- 60 C/C++ pull requests from GitHub
  - 20 adopted from a previous study
  - 40 from the top 60 starred C/C++ projects
- **Author images**: Relabel the author information
  - Human: man, woman
    - Chicago Face Database (CFD)
Experimental Design: Code Review Tasks

- 60 C/C++ pull requests from GitHub
  - 20 adopted from a previous study
  - 40 from the top 60 starred C/C++ projects
- Author images: **Relabel the author information**
  - Human: man, woman
    - Chicago Face Database (CFD)
  - Machine (APR Tools)
Experimental Design: Code Review Tasks

- 60 C/C++ pull requests from GitHub
- Author images: Relabel the author information
- Construction of code review stimuli

60 Pull Requests + 20 Men + 20 Women + 20 Machine = 60 Stimuli
Experimental Design: Code Review Tasks

- 60 C/C++ pull requests from GitHub
- Author images: Relabel the author information
- Construction of code review stimuli: two versions

60 Pull Requests + 20 Men + 20 Machine = 60 Stimuli
Experimental Design: Code Review Tasks

- 60 C/C++ pull requests from GitHub
- Author images: Relabel the author information
- Construction of code review stimuli

60 Stimuli: V1

60 Stimuli: V2

Delete the equal mark in case the array is like \{x,x,x\ldots(n),y,y,y\ldots(n+1)\}
Experimental Design: Code Review Tasks

- 60 C/C++ pull requests from GitHub
- Author images: Relabel the author information
- Construction of code review stimuli

Please wait for the next pull request submitted by this programmer:

Name: [Name]
Affiliation: [Affiliation]
Title: [Title]

Next pull request is loading ...

Delete the equal mark in case the array is like \{x,x,x...\(n\),y,y,y,...\(n+1\)\}

Owner: [Owner]

Accept  Reject

Yu Huang @ FSE 2020
Experimental Design

Social desirability bias
“This study is to investigate how software developers conduct code reviews.”

“All of the pull requests are from real-world software projects and development teams.”

“Some of the pull requests are generated by computer programs.”
“Sorry.”

“Actually, this study is to check biases on genders and identities of authors in code review.”

“All of the pull requests are made by human developers. None is generated by machines.”

“All the profile pictures are randomly assigned.”
Experimental Design: Recruitment

- 37 participants
  - Native English speakers
  - Left-handed

<table>
<thead>
<tr>
<th>Demographic</th>
<th>Number of Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
</tr>
<tr>
<td>Men</td>
<td>21</td>
</tr>
<tr>
<td>Women</td>
<td>16</td>
</tr>
<tr>
<td>Undergraduate</td>
<td>26</td>
</tr>
<tr>
<td>Graduate</td>
<td>11</td>
</tr>
</tbody>
</table>
Experimental Design: Post Survey

- How would you **compare** the machine-generated code changes (i.e., by automated repair tools) with the human-generated changes?
- Do you think there are any difference between code written by men and women?
Research Questions

- **RQ1**: How do the identities of code reviewers and authors change or bias the code review process behaviorally?
- **RQ2**: Can we differentiate the gender identities of code reviewers based on their visual attention patterns?
- **RQ3**: Can we classify the gender identities of code reviewers based on patterns of brain activity?
- **RQ4**: How do self-reports of the role of identity in code review align with reality?
Results:

- **RQ1**: How do the identities of code reviewers and authors change or bias the code review process behaviorally?
  - Behaviorally, men and women conduct code reviews differently
    - LMM, statistical tests
  - All participants spend *less time* evaluating the Pull Requests of women $(p<0.01)$
  - All participants are *less likely to accept* the Pull Requests of machines $(p<0.05)$
  - *Women reviewers* spent *less time* on all Pull Requests than men $(p<0.0001)$
Results

• **RQ2:** Can we differentiate the gender identities of code reviewers based on their visual attention patterns?

  ○ Men and women participants employ *different high-level problem-solving strategies* in code review.

  ● Men fixated more frequently \((p<0.001)\), while women spent significantly more time analyzing Pull Requests messages and author pictures.

  ![Eye-tracking: Fixation Time Distribution](chart)

**Eye-tracking: Fixation Time Distribution**

- **Author Pic:** Men: 8.60%, Women: 12.60%
- **Code:** Men: 62.50%, Women: 57.30%
- **PR Message:** Men: 12.50%, Women: 25.94%
- **Indicator Image:** Men: 3.59%, Women: 16.40%
• **RQ2:** Can we differentiate the gender identities of code reviewers based on their visual attention patterns?

Results

(a) A stimulus with a machine author
(b) A stimulus with a woman author
(c) A stimulus with a man author
Results

● **RQ3:** Can we classify the gender identities of code reviewers based on patterns of brain activity?
  ○ Relative to women reviewers, *men* show *less consistent differences* in their responses to woman- vs. man-authored Pull Requests.
  ● Gaussian Process Classification
    ○ It is possible to *distinguish* women and men conducting code review at a neurological level (BAC=68.59%, p=0.016).
Results

- **RQ4**: How do **self-reports** of the role of identity in code review **align with reality**?
  - Although humans exhibit biases in their acceptance rates of identical code labeled as written by women vs men vs. machines, participant self-reports **only acknowledge the bias against machines** *(23 : 8)* **but do not acknowledge a gender bias**.
  - When Pull Request author information changes, participants report **seeing quality differences where none exist** *(reported: machines-generated code has lower quality)*.

  "**Machine-generated changes are IMO less readable, a little worse in quality, capable in fewer scopes**"
Summary

- We present a **controlled experiment** using both **medical imaging** and **eye-tracking** to investigate **biases and differences** in code review.
  - Genders, humans, machines
- We find **universal biases** in how all participants treat code reviews as a function of the reviewers’ gender and apparent author:
  - Behavioral difference
  - Visual difference
  - Neurological difference
- We find participants’ **self-reported perception** of decision making in code review **do not align** with the objective observations.
  - Bias against machines exists
  - Do not realize the existence of difference on gender
Motivation

- Code review is critical for software development
  - Systematic inspection, analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be 60%-65%

Decoding the representation of code in the brain:
An fMRI study of code review and expertise

Benjamin Floyd  
University of Virginia  
bef2cj@virginia.edu

Tyler Santander  
University of Virginia  
ts7ar@virginia.edu

Westley Weimer  
University of Virginia  
weimer@virginia.edu
Motivation

- Code review is critical for software development
  - Systematic inspection, analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be 60%-65%
• **RQ1**: How do the identities of code reviewers and authors change or bias the code review process? **Behavioral Difference**

- Behaviorally, men and women conduct code reviews differently

<table>
<thead>
<tr>
<th>Author Label</th>
<th>Woman</th>
<th>Man</th>
<th>Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Time (s)</td>
<td>20.8</td>
<td>21.7</td>
<td>21.7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reviewer’s Gender</th>
<th>Woman</th>
<th>Man</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Time (s)</td>
<td>20.5</td>
<td>22.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Author Label</th>
<th>Woman</th>
<th>Man</th>
<th>Machine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acceptance Rate</td>
<td>84.36%</td>
<td>79.68%</td>
<td>78.03%</td>
</tr>
</tbody>
</table>
Results

● **RQ2: Can we classify the gender identities of code reviewers based on patterns of brain activity?** **Neurological Difference**
  ○ Relative to women reviewers, men show less consistent differences in their responses to woman- vs. man-authored Pull Requests.
  ● Gaussian Process Classification
  ○ It is possible to distinguish women and men conducting code review at a neurological level (BAC=68.59%, p=0.016). Men and women conduct code reviews differently in terms of associated cognitive processes and patterns of neural activation.