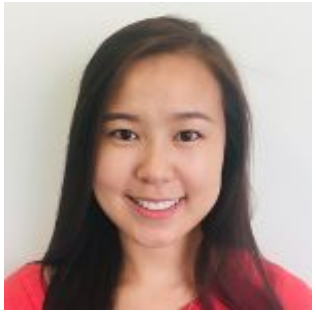
# Biases and Differences in Code Review Using Medical Imaging and Eye-Tracking: Genders, Humans, and Machines

**Yu Huang**[1], Kevin Leach[1], Zohreh Sharafi[1], Nicolas McKay[1], Tyler Santander[2], Westley Weimer[1]

[1]University of Michigan, Ann Arbor
[2]University of California, Santa Barbara

COLLEGE OF ENGINEERING
COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

# Thank You to the Collaborators!



Yu Huang

Dr. Kevin Leach

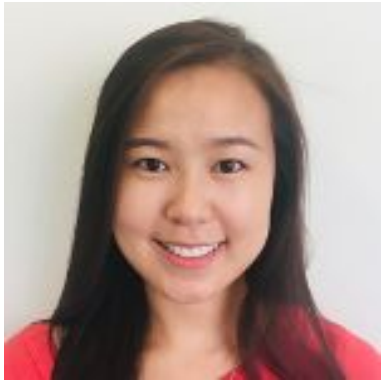
Dr. Zohreh Sharafi

Nicholas McKay

Dr. Tyler Santander

Dr. Westley Weimer

# On the Academic Job Market!



**Yu Huang**

**SE & Human Factors**

yhhy@umich.edu

**Dr. Kevin Leach**

**SE & Security**

kjleach@umich.edu

**Dr. Zohreh 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%**

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

2 algorithms/cpp/majorityElement/majorityElement.cpp

```
@@ -32,7 +32,7 @@ int majorityElement(vector<int> &num) {
32 32                cnt++;
33 33            }else{
34 34                majority == num[i] ? cnt++ : cnt --;
35    -              if (cnt >= num.size()/2) return majority;
   35 +              if (cnt > num.size()/2) return majority;
36 36            }
37 37        }
38 38        return majority;
```

# 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%**

**Code changes**

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

algorithms/cpp/majorityElement/majorityElement.cpp

```
@@ -32,7 +32,7 @@ int majorityElement(vector<int> &num) {
                cnt++;
            }else{
                majority == num[i] ? cnt++ : cnt --;
-               if (cnt >= num.size()/2) return majority;
+               if (cnt > num.size()/2) return majority;
            }
        }
        return majority;
```

# 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**

**Code changes**

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

2 algorithms/cpp/majorityElement/majorityElement.cpp

```
@@ -32,7 +32,7 @@ int majorityElement(vector<int> &num) {
                cnt++;
            }else{
                majority == num[i] ? cnt++ : cnt --;
-               if (cnt >= num.size()/2) return majority;
+               if (cnt > num.size()/2) return majority;
            }
        }
        return majority;
```

# 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%**

Beyond the Code Itself:
How Programmers *Really* Look at Pull Requests


Denae Ford, Mahnaz Behroozi
*North Carolina State University*
Raleigh, NC, USA
{dford3, mbehroo}@ncsu.edu

Alexander Serebrenik
*Eindhoven University of Technology*
Eindhoven, The Netherlands
a.serebrenik@tue.nl

Chris Parnin
*North Carolina State University*
Raleigh, NC, USA
cjparnin@ncsu.edu


# Motivation

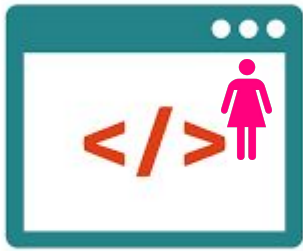
- **Code review is *critical* for software development**
  - ***Systematic*** inspection**,** analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be **60%-65%**

Beyond the Code Itself:
How Programmers *Really* Look at Pull Requests


Denae Ford, Mahnaz Behroozi
*North Carolina State University*
Raleigh, NC, USA
{dford3, mbehroo}@ncsu.edu

Alexander Serebrenik
*Eindhoven University of Technology*
Eindhoven, The Netherlands
a.serebrenik@tue.nl

Chris Parnin
*North Carolina State University*
Raleigh, NC, USA
cjparnin@ncsu.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%**

Beyond the Code Itself:
How Programmers *Really* Look at Pull Requests

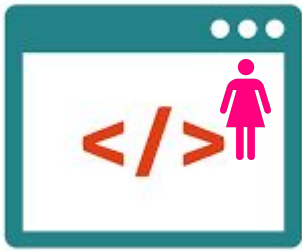
Denae Ford, Mahnaz Behroozi
*North Carolina State University*
Raleigh, NC, USA
{dford3, mbehroo}@ncsu.edu

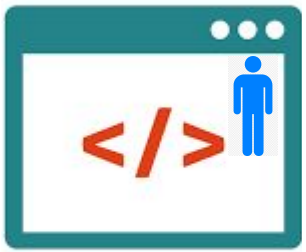
Alexander Serebrenik
*Eindhoven University of Technology*
Eindhoven, The Netherlands
a.serebrenik@tue.nl

Chris Parnin
*North Carolina State University*
Raleigh, NC, USA
cjparnin@ncsu.edu

**Gender differences and bias in open source: pull request acceptance of women versus men**

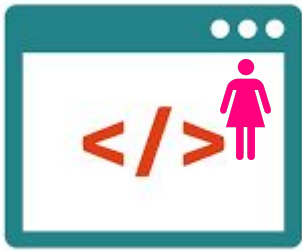
Josh Terrell[1], Andrew Kofink[2], Justin Middleton[2], Clarissa Rainear[2], Emerson Murphy-Hill[2], Chris Parnin[2] and Jon Stallings[3]

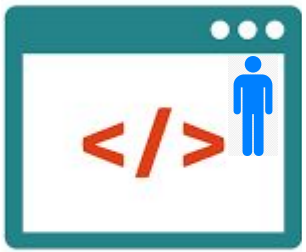




# Motivation

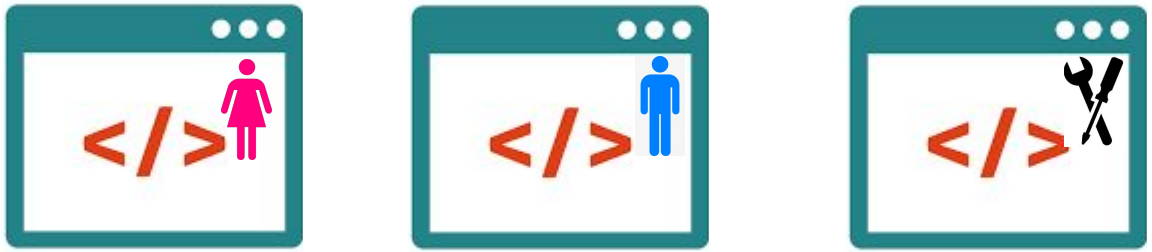
- **Code review is *critical* for software development**
  - ***Systematic*** inspection**,** analysis, evaluation, and revision of code.
  - Latent defect discovery rate of formal code review can be **60%-65%**



# 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%**


Current challenges in auto

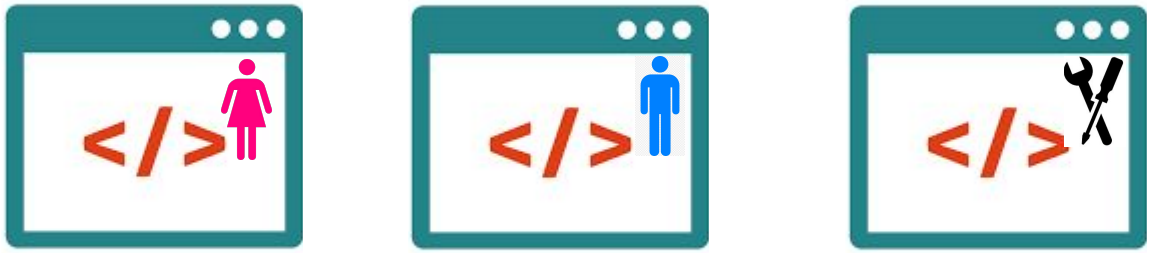
Claire Le Goues · Stephanie Forrest

Trust in Automated Software Repair

The Effects of Repair Source, Transparency, and Programmer Experience on Perceived Trustworthiness and Trust

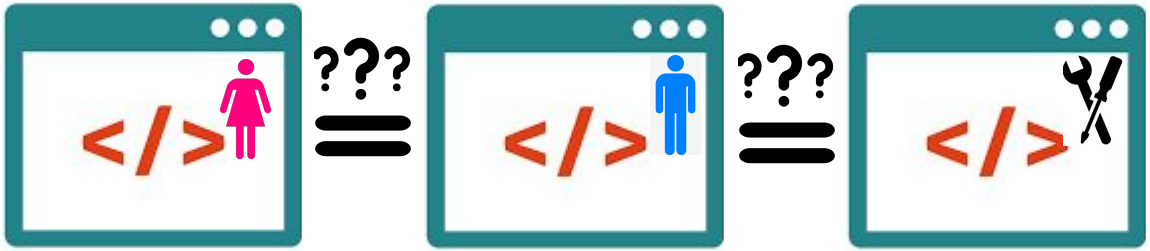
Authors | Authors and affiliations

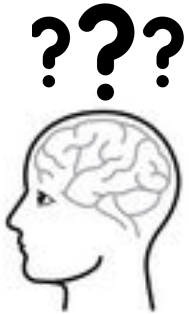
Tyler J. Ryan, Gene M. Alarcon, Charles Walter, Rose Gamble, Sarah A. Jessup, August Capiola, Marc D. Pfahler



# High-level Question

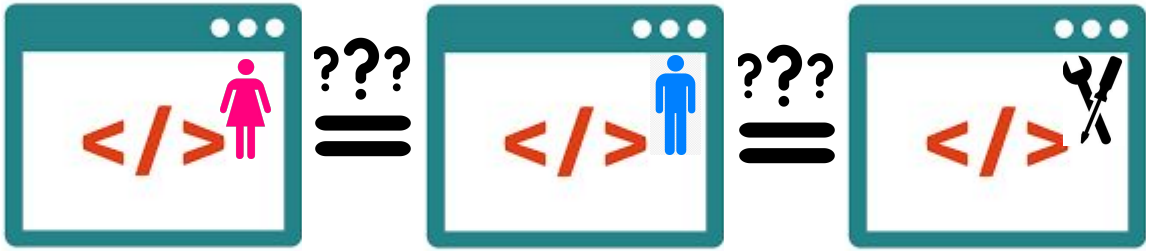
- Is there **bias** on **gender** and **identities** in code review? How do we characterize the bias?



# 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**

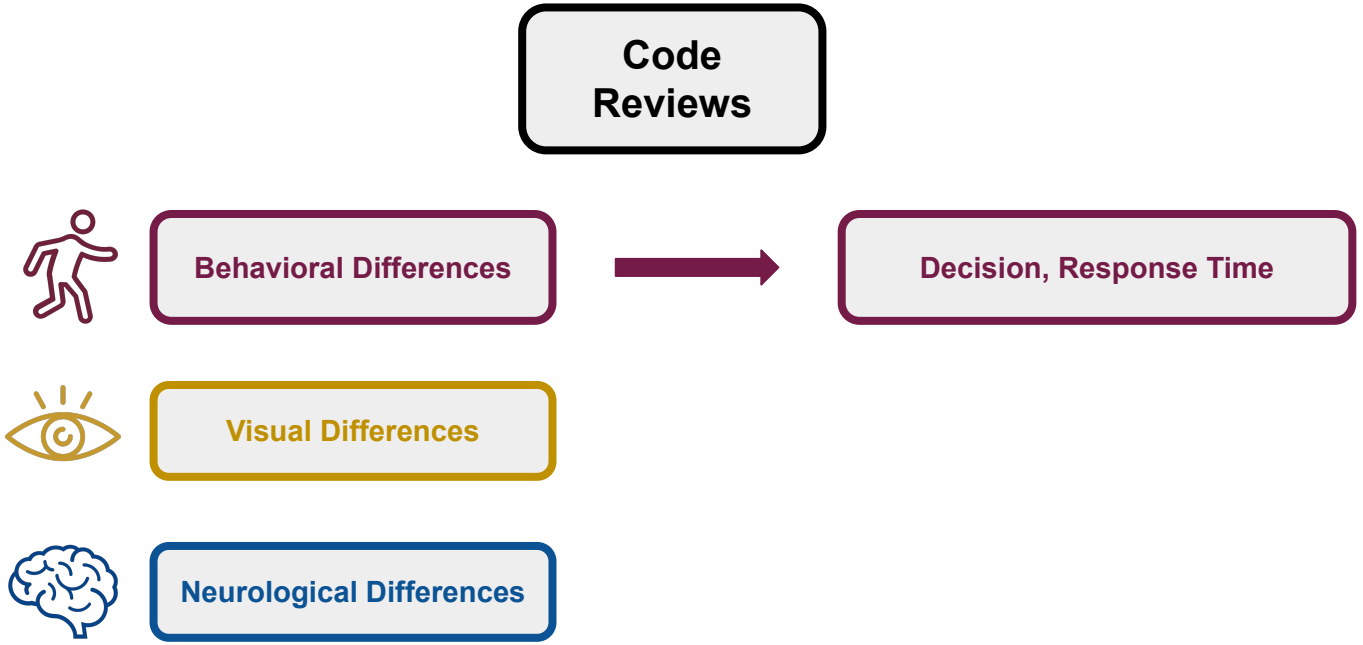
**Neurological Differences**

# Outline

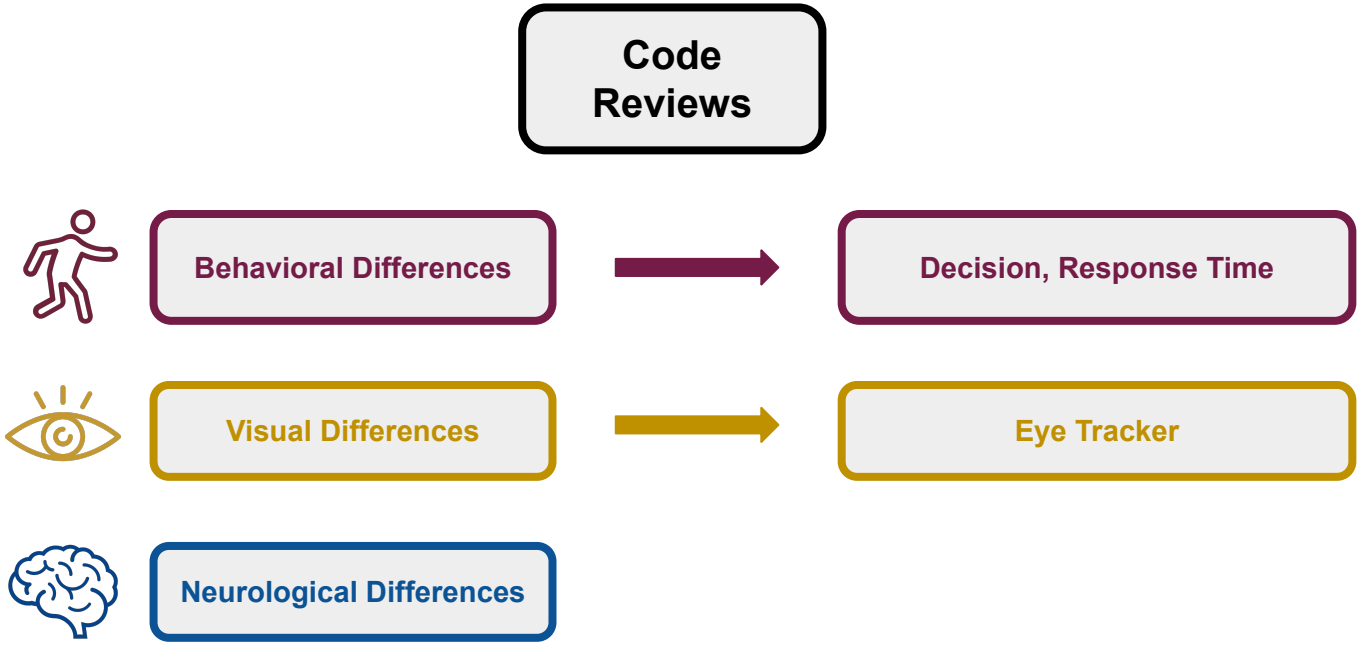
- Motivation
- High-level question
- **Experimental design**
- **Results**
- **Conclusions**

# Experimental Design: Code Review Tasks

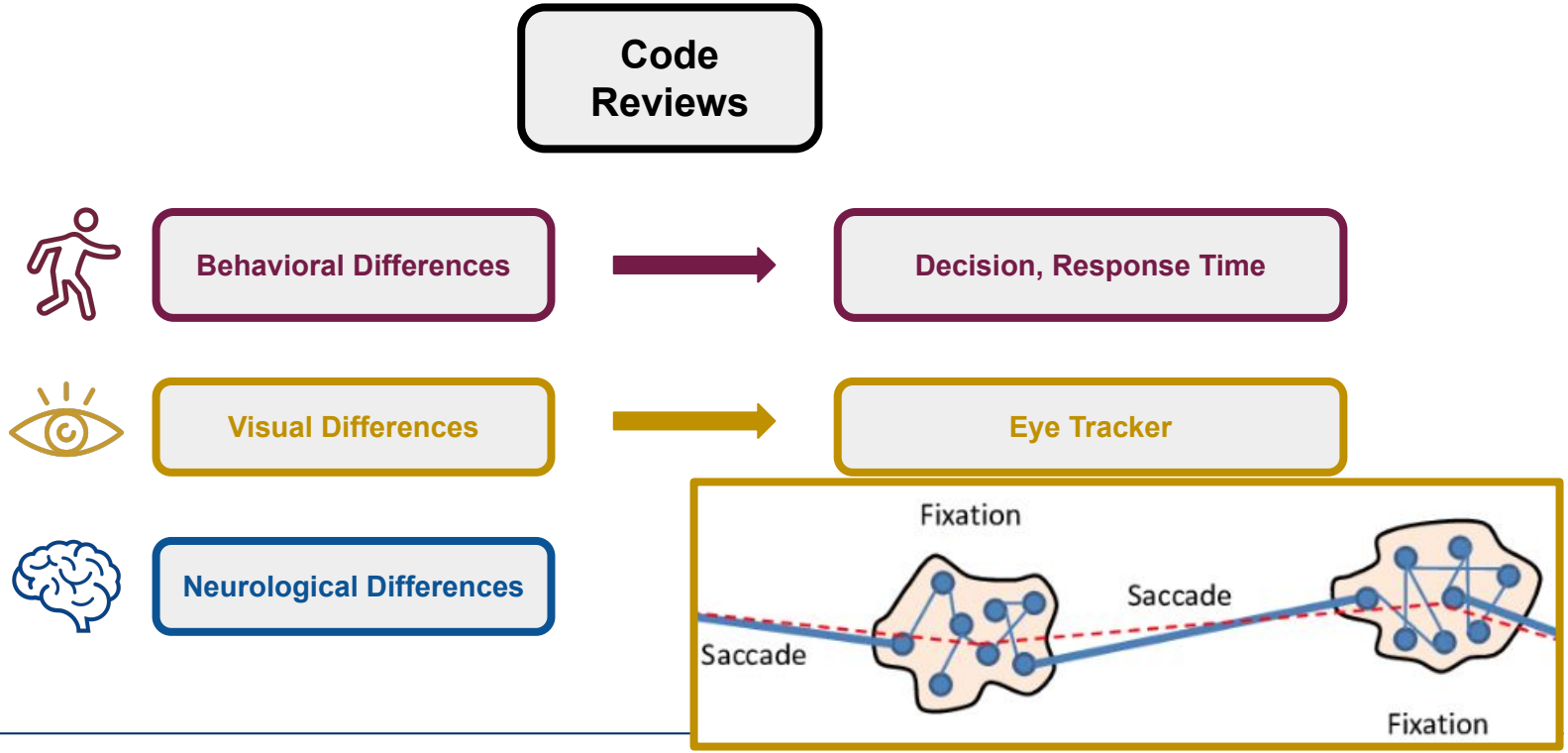
# Experimental Design: Code Review Tasks



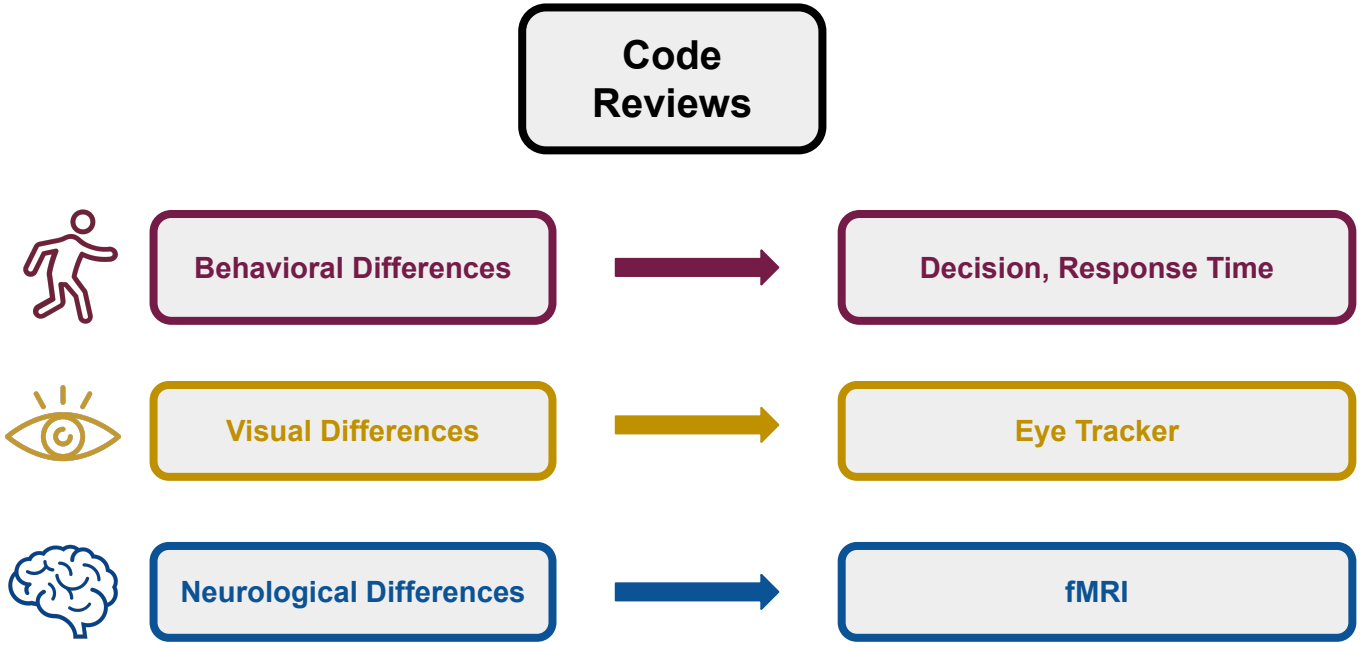
# Experimental Design: Code Review Tasks



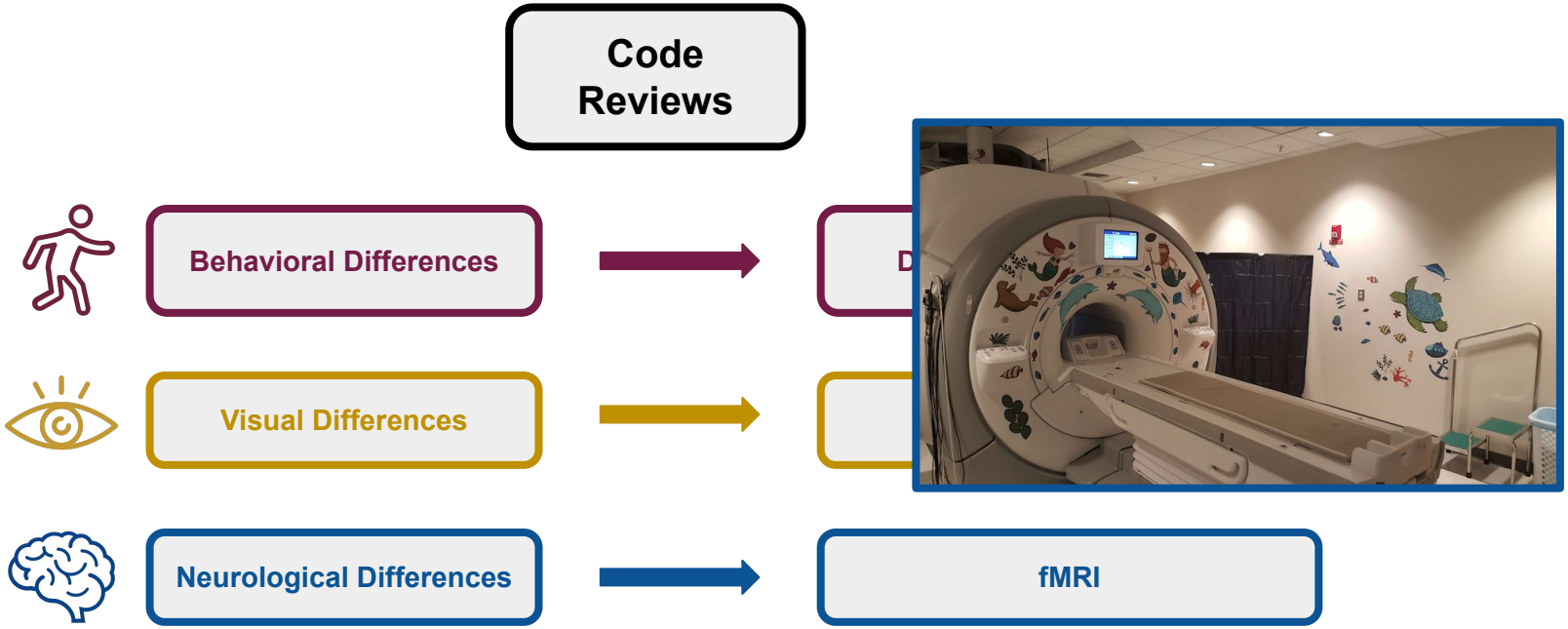
# Experimental Design: Code Review Tasks



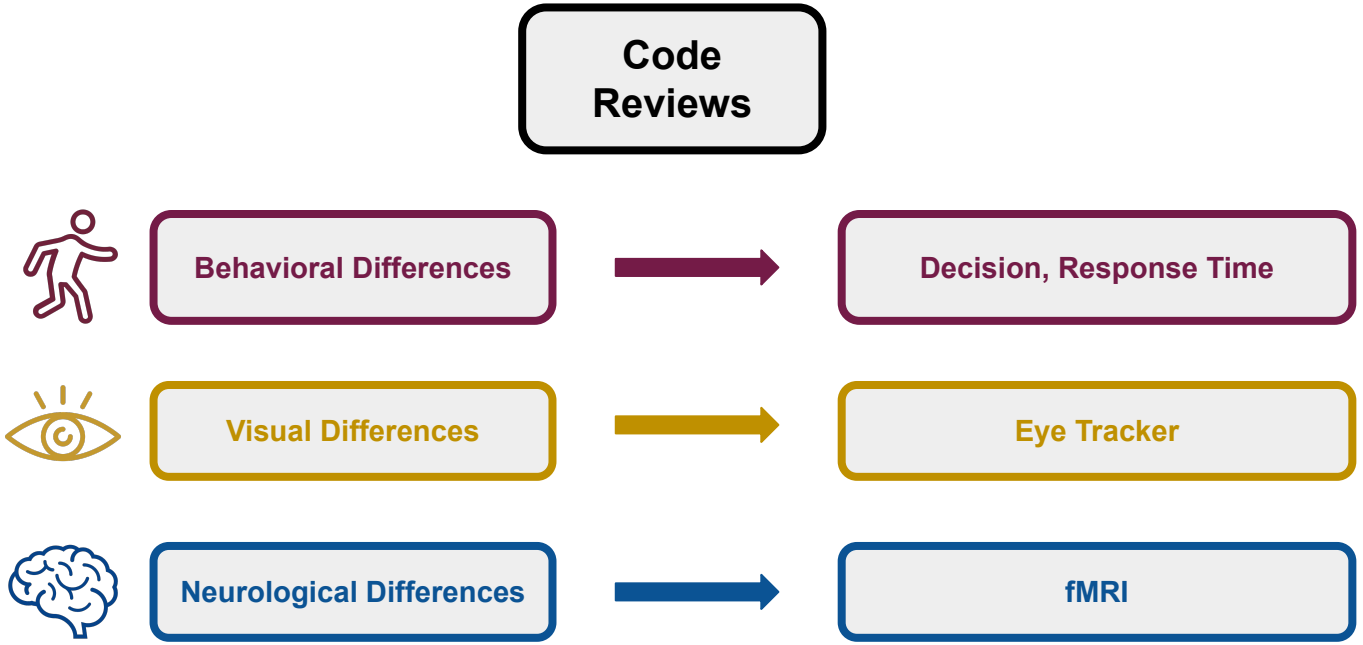
# Experimental Design: Code Review Tasks



# Experimental Design: Code Review Tasks



# Experimental Design: Code Review Tasks



# 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

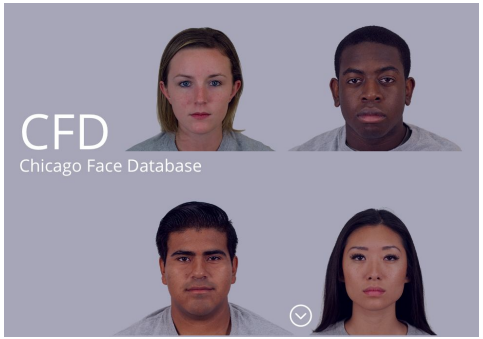

**Delete the equal mark in case the array is like {x,x,x...(n),y,y,y,y...(n+1)}**

2 algorithms/cpp/majorityElement/majorityElement.cpp

```
@@ -32,7 +32,7 @@ int majorityElement(vector<int> &num) {
                cnt++;
            }else{
                majority == num[i] ? cnt++ : cnt --;
-               if (cnt >= num.size()/2) return majority;
+               if (cnt > num.size()/2) return majority;
            }
        }
        return majority;
```

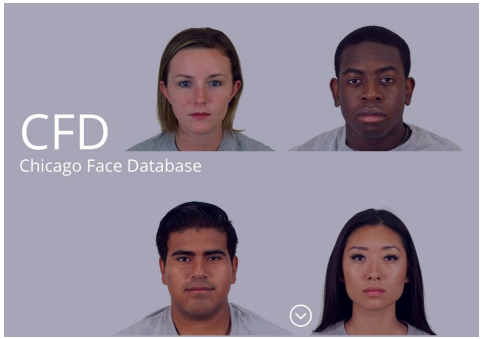
# Experimental Design: Code Review Tasks

- **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

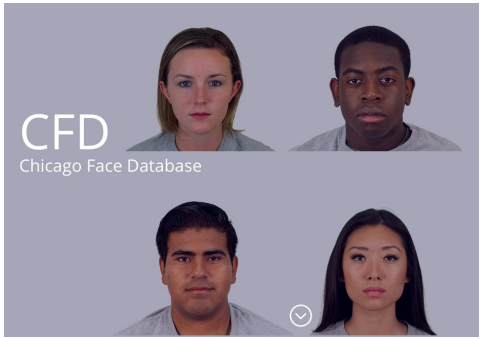
# Experimental Design: Code Review Tasks

- **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)



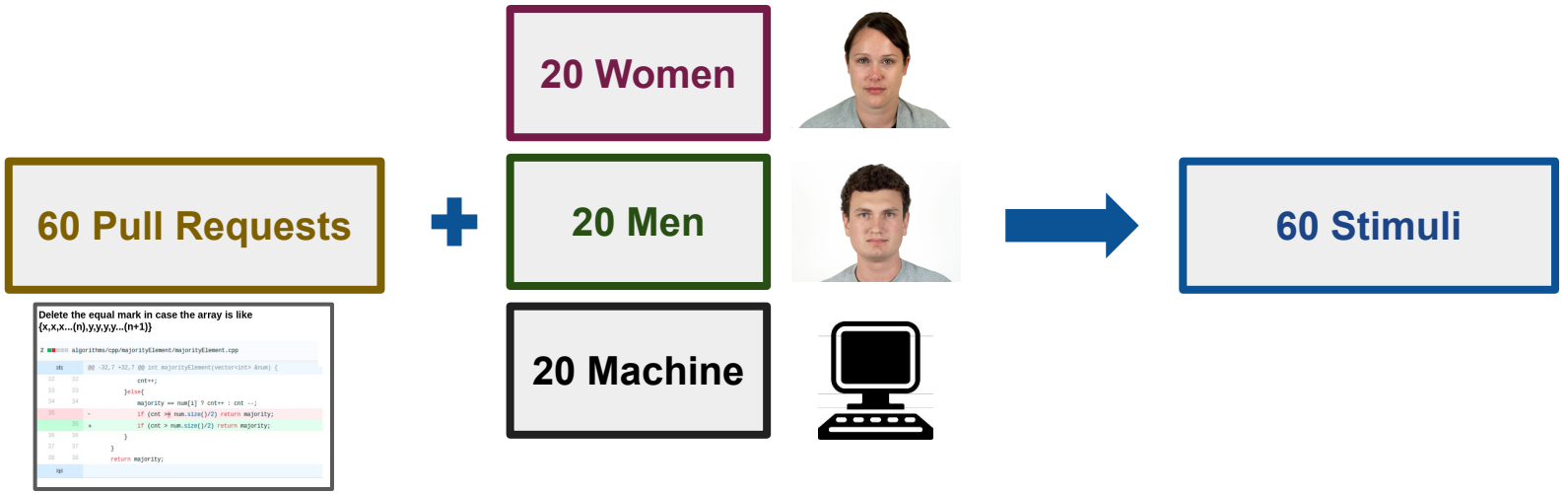

# Experimental Design: Code Review Tasks

- **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)

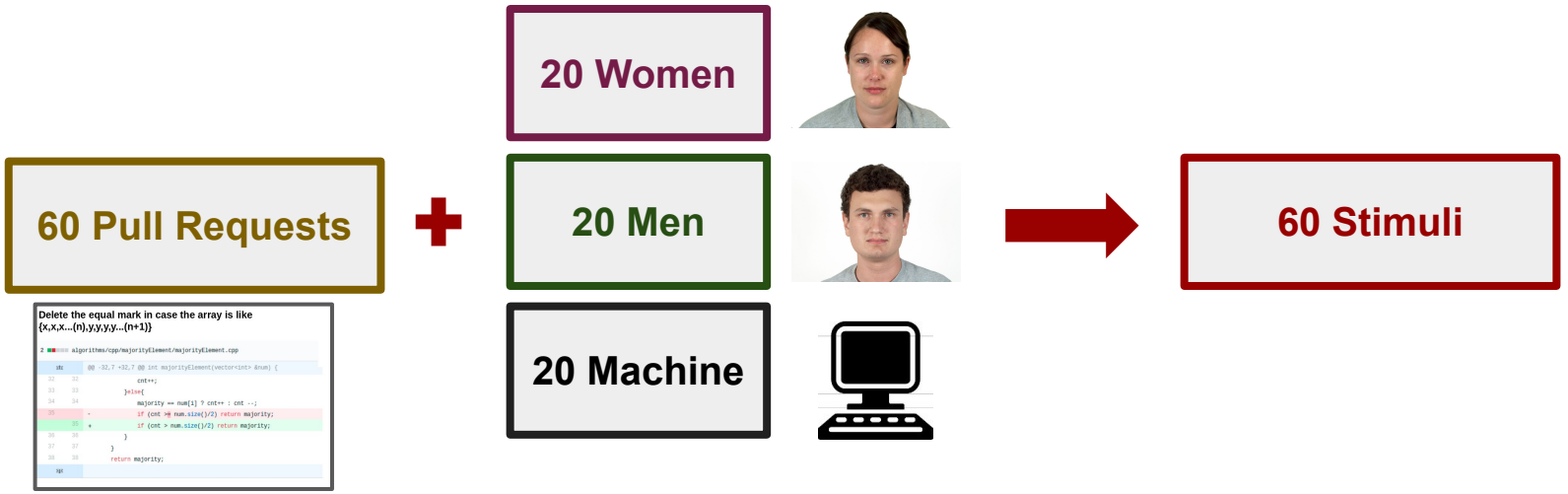

# Experimental Design: Code Review Tasks

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



# 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*

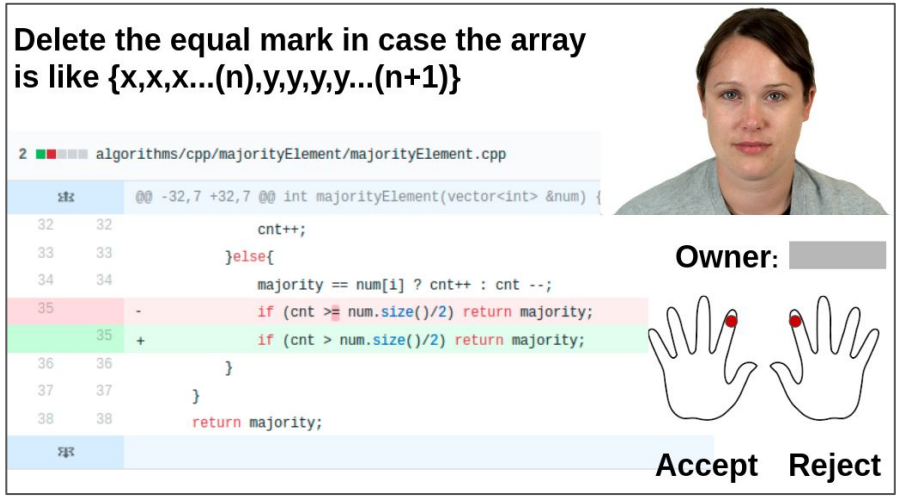


# Experimental Design: Code Review Tasks

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

60 Stimuli: V1

60 Stimuli: V2



# 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:
Affiliation:
Title:

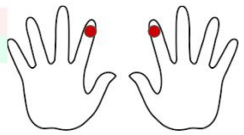
Next pull request is loading ...

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

2 algorithms/cpp/majorityElement/majorityElement.cpp

```
@@ -32,7 +32,7 @@ int majorityElement(vector<int> &num) {
32 32            cnt++;
33 33        }else{
34 34            majority == num[i] ? cnt++ : cnt --;
35    -          if (cnt >= num.size()/2) return majority;
   35 +          if (cnt > num.size()/2) return majority;
36 36        }
37 37    }
38 38    return majority;
```

Owner:

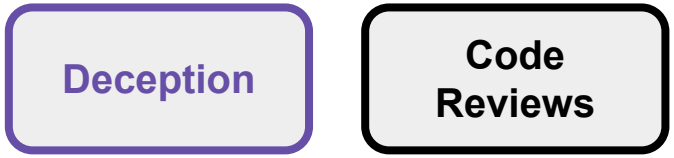


Accept Reject

# Experimental Design


**Social desirability bias**

# Experimental Design: Deception




*"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."*

# Experimental Design: Deception


*"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

**Recruitment** | **Deception** | **Code Reviews** | **Debriefing**

- **37 participants**
  - Native English speakers
  - Left-handed

| Demographic | Number of Participants | | |
|---|---|---|---|
| | Total | Version I | Version II |
| Men | 21 | 11 | 10 |
| Women | 16 | 7 | 9 |
| Undergraduate | 26 | 11 | 15 |
| Graduate | 11 | 7 | 4 |

# 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

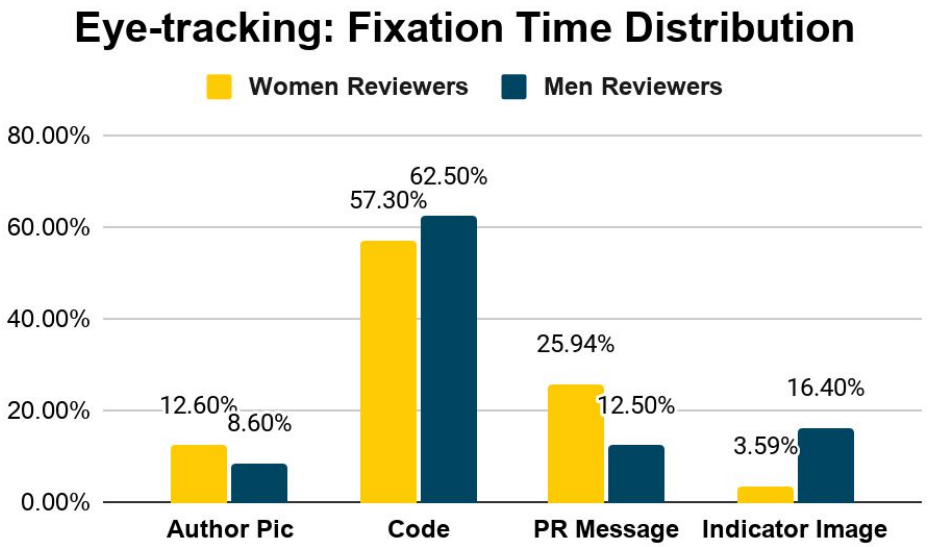
- **RQ1:** How do the identities of code **reviewers** and **authors** change or bias the code review process behaviorally?
- **RQ2:** Can we differentiate the gender identities of code **reviewers** based on their **visual attention patterns**?
- **RQ3:** Can we classify the gender identities of code **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 ($p<0.0001$)

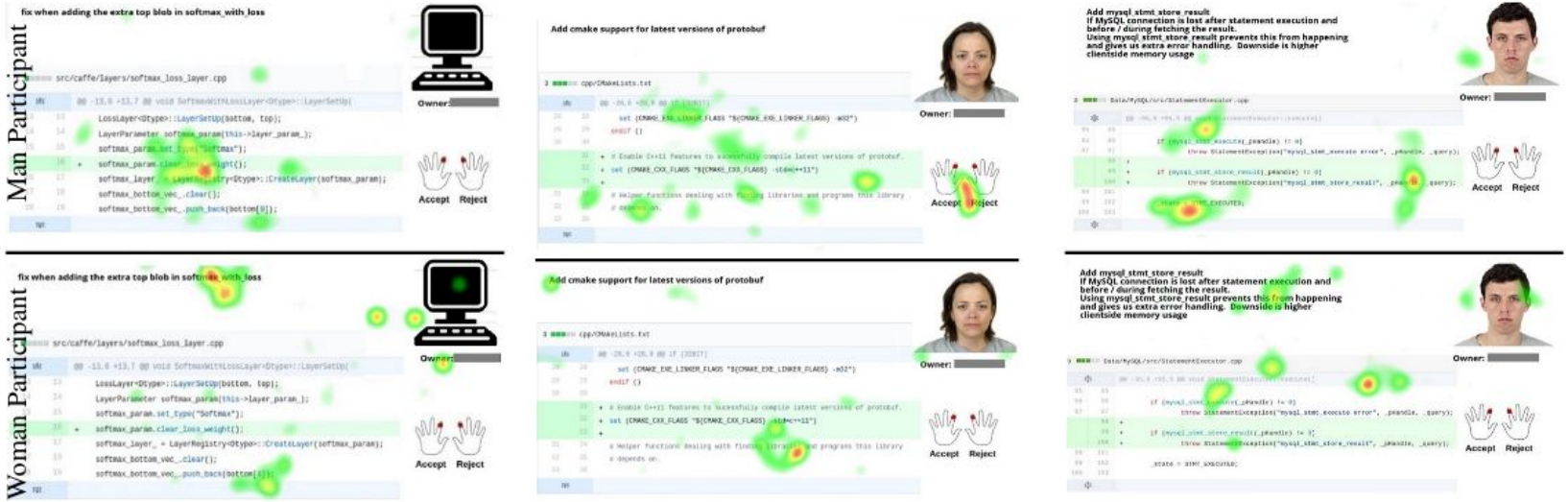
# Results

- **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.



# Results

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




(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

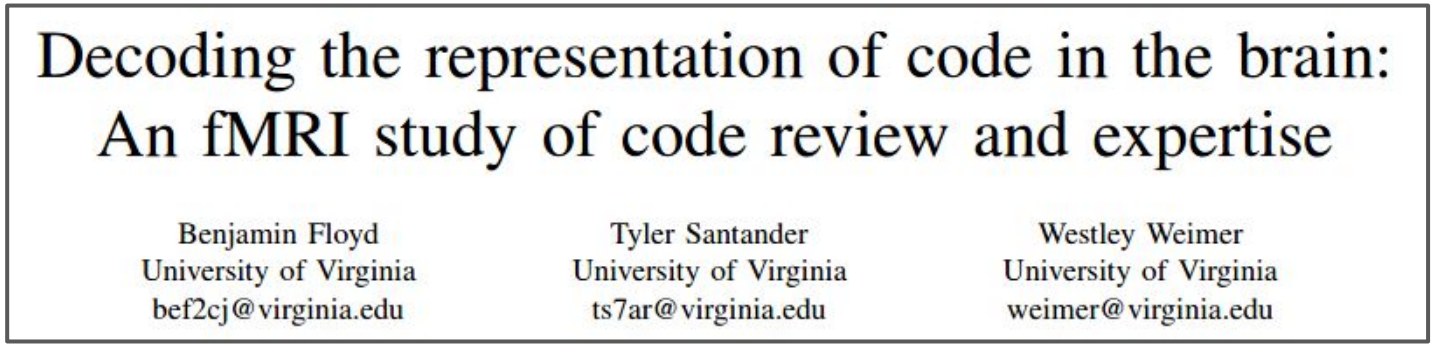



- We present a **controlled experiment** using both medical imaging and eye-tracking to investigate **biases and differences** in code review.
  - Genders, humans, machines
- We find **universal biases** in how all participants treat code reviews as a function of the **reviewers' gender** and **apparent 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

# Bonus

# 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

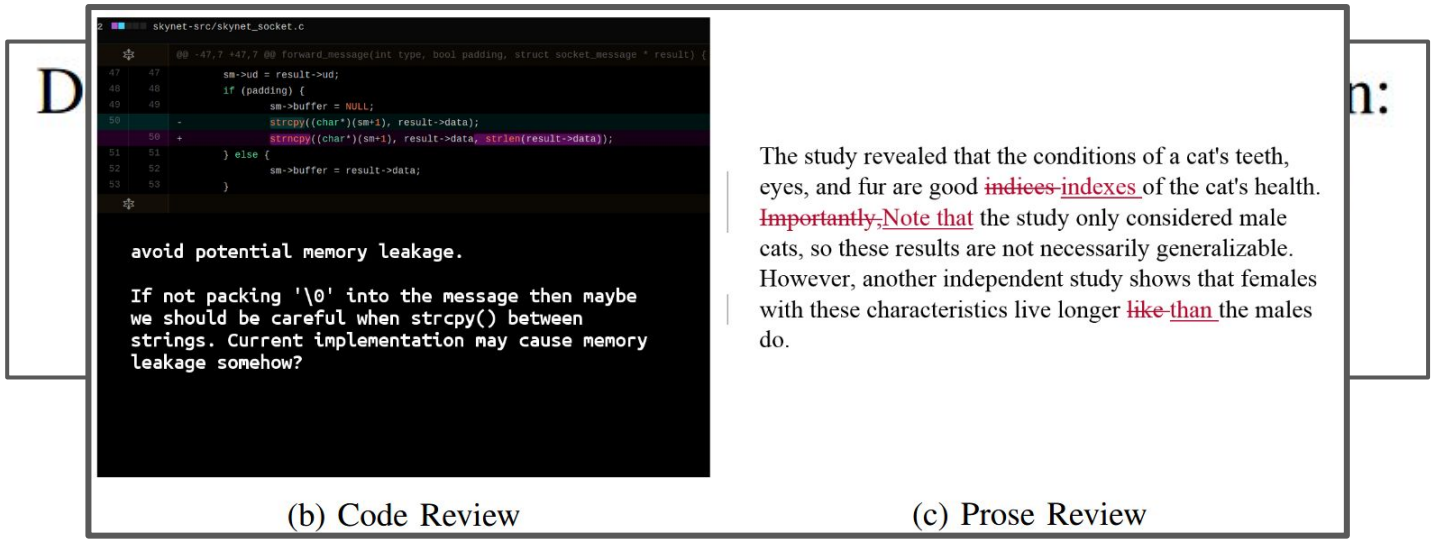

Benjamin Floyd
University of Virginia
bef2cj@virginia.edu

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%




(b) Code Review (c) Prose Review

# Results

- **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**
  - LMM, statistical tests

| Author Label | Woman | Man | Machine |
|---|---|---|---|
| Response Time (s) | 20.8 | 21.7 | 21.7 |

| Reviewer's Gender | Woman | Man |
|---|---|---|
| Response Time (s) | 20.5 | 22.1 |

| Author Label | Woman | Man | Machine |
|---|---|---|---|
| Acceptance Rate | 84.36% | 79.68% | 78.03% |

# 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