Do Users Act Equitably? Understanding User Bias Through a Large In-person Study

Yang Liu, Heather Moses, Mark Sternefeld, Samuel Malachowsky and Daniel E. Krutz
Department of Software Engineering
Rochester Institute of Technology, Rochester, NY, USA
Email: {yl4070, hlm8500, mfs5944, samvse, dxkvse}@rit.edu

***Abstract*—Inequitable software is a common problem. Bias may be caused by developers, or even software users. As a society, it is crucial that we understand and identify the causes and implications of software bias from both users and the software itself. To address the problems of inequitable software, it is essential that we inform and motivate the next generation of software developers regarding bias and its adverse impacts. However, research shows that there is a lack of easily adoptable ethics-focused educational material to support this effort.**

To address the problem of inequitable software, we created an easily adoptable, self-contained experiential activity that is designed to foster student interest in software ethics, with a specific emphasis on AI/ML bias. This activity involves participants selecting fictitious teammates based solely on their appearance. The participant then experiences bias either against themselves or a teammate by the activity's fictitious AI. The created lab was then utilized in this study involving 173 real-world users (age 18-51+) to better understand user bias.

The primary findings of our study include: I) Participants from minority ethnic groups have stronger feeling regarding being impacted by inequitable software/AI, II) Participants with higher interest in AI/ML have a higher belief for the priority of unbiased software, III) Users do not act in an equitable manner, as avatars with 'dark' skin color are less likely to be selected, and IV) Participants from different demographic groups exhibit similar behavior bias. The created experiential lab activity may be executed using only a browser and internet connection, and is publicly available on our project website: https://all.rit.edu.

***Index Terms*—Accessibility Education, Computing Education, Computing Accessibility**

General Abstract

Inequitable software is a significant problem in today's society. Unfortunately, there is a lack of easily adoptable experiential educational materials that educators and practitioners can use to demonstrate and understand the impacts of inequitable software. To address this challenge, we developed a hosted educational experiential activity to provide a mechanism for instructors, students, and practitioners to experience the adverse impacts of bias software firsthand. We then utilized this activity as a basis for a large in-person study involving 173 real-world users to better understand user bias.

Our hosted, experiential activity demonstrates the adverse impacts of inequitable software and bias. This is accomplished by inflicting bias against the participant and fictitious teammates in a simple tic-tac-toe game. The activity may be adopted using only a web browser and internet connection. The activity is available on the project website: https://all.rit.edu.

To better understand user bias, the created experiential activity was utilized in a large in-person study. The primary findings of this study are that: participants from minority ethnicity demographics have stronger feeling towards inequitable software/AI; participants with higher interest in machine learning and AI have higher beliefs in the prioritization of unbiased software, and that users do not act in an equitable manner – choosing teammates of color at a disproportionately low rate.

This work benefits educators by providing an easily adoptable experiential educational activity that they may use to demonstrate the adverse impacts of inequitable software in the classroom. The activity may also be used to foster discussions pertaining to this foundational and essential topic. Researchers will benefit from this work through an increased understanding of inequitable decisions made by users.

I. Introduction

Bias comes in many shapes and forms, and is present in numerous computing systems [19, 44]. Bias can stem from computers (AI/ML) or directly from humans [46, 62]. Better understanding the causes and impacts of bias in the real-world is an important step for properly addressing this issue. As a society, we clearly need to do all that we can to combat inequity and bias in software in every possible manner.

Unfortunately, there is a lack of robust and easily adoptable educational materials concerning AI ethics that can be used to demonstrate the adverse impacts of bias to students [33, 51]. Therefore, educators are often restricted from including ethically-focused AI/ML topics in their curriculum. This issue is especially prevalent at many smaller and/or less well-funded institutions (that frequently serve underrepresented groups), which impedes the inclusion of these essential topics in their curriculum [29, 31]. To address these issues, we: I) Created an easily adoptable, ethics-focused AI experiential educational lab to support the inclusion of this vital topic in a myriad of computing and non-computing courses, and II) Utilized this lab as a foundational activity to better understand user bias.

Our web hosted experiential educational lab [1, 32] has users select fictitious teammates, using only avatars representing the appearance of themselves and their teammates for a tic-tac-toe game. These avatars represent a diverse range of teammates (*e.g.,* gender, skin colors, etc.). During the activity, participants will experience either bias against themselves or their teammates due to a fabricated AI bias in the system

against whatever demographics were chosen by the participant (*e.g.,* a bias against people of color, women, etc.). At the conclusion of the activity, participants are presented with various issues of bias that they incorporated into the application (*i.e.,* selecting a disproportionate number of teammates from a specific gender or skin color). This lab can be used in many settings, ranging from outreach events to conventional computing and non-computing focused classrooms. This lab may be especially useful for fostering discussions about inclusive and equitable AI/ML. This hosted educational lab [1, 32] supports easy adoption, requiring only a browser for usage.

The created educational lab served as a foundational activity to provide insight into our study to better understand user bias. We accomplished this by conducting a large in-person study involving 173 real-world participants, whose age ranged from 18 - $\geq$ 51, to provide constructive insights for developers, educators, and researchers. During this study, community participants from diverse backgrounds and ages were informed that they were playing a tic-tac-toe game against a fictitious AI. Participants first selected an avatar that best represented their appearance, and then selected three team members (based entirely on the avatar appearances) under the guise that these team members were other human players.

Our principle observations include: I) Participants from minority ethnic groups have stronger feeling regarding being impacted by inequitable software/AI, II) Participants with higher interest in AI/ML have a higher belief for the priority of unbiased software, III) Users do not act in an equitable manner, as avatars with 'dark' skin color are less likely to be selected, and IV) Participants from different demographic groups exhibit similar behavior bias.

To summarize, this work makes the following contributions:

- **Public bias-oriented experiential educational activity:** Our experiential educational activity demonstrates the importance of creating equitable software to participants. This hosted activity requires only a browser for usage[1].
- **Experimental findings:** We found that participants did not act equitably, choosing teammates with the 'black' skin color at a disproportionately low rate. We also found that participants from minority ethnic groups feel that they are more likely to be impacted by inequitable software in the real-world.

The rest of the paper is organized as follows: Section II presents our created experiential education lab, and Section III discusses the design of our study. Evaluation results are provided in Section IV and Section V discusses the findings of our research. Section VI presents related works, and Section VII provides a conclusion.

II. Developed Experiential Intervention Lab

In the following sections, we will describe the structure of the self-contained publicly provided lab activity [1, 32], along with the lab itself. The provided instructional lab activity differs from that used in our experiment (Section III) in that our experiment did not contain foundational instructional or assessment materials such as reading, quiz, etc. due to the brevity of the intervention.

[1]Note: Lab link removed due to double-anonymous requirements

A. Lab Structure

The developed experiential lab contains the components shown in Table I. These components are systematically designed to support activity that is I) Informal, II) Interesting for the participant, and III) Easy to adopt with minimal resource requirements due to its hosted nature.

Component	Description
Information on ML/AI Bias	Background information on bias and other ethical concerns in AI/ML. This will include discussion and thought-provoking material.
Lab instructions	Information on how to complete the lab.
Lecture slides	Example lecture slides that an instructor may use to present the material. The slides will be in ppt format, so instructors can still alter them as they'd like.
Video presentation of lecture slides	ADA-compliant YouTube video of project member presenting the lecture. This will support students conducting the lab on their own, or the flipped classroom environment.
Experiential activity	The participant utilizes the hosted, experiential activity that addresses bias in AI/ML. This is essentially the activity used in the experiment (Section III).
Video of lab being conducted	ADA-compliant YouTube video of the lab being conducted.
Quiz	An example quiz is provided to instructors to evaluate their students at the conclusion of the activity if they desire to do so.

TABLE I: Lab Components

B. Lab Learning Objectives

Learning Objectives (LO) - After completion of the lab, participants should be able to:

LO1: Recognize different ethical challenges and bias in an intelligent system (Comprehension).

LO2: Diagnose ethical implications of choices made by AI (Syntheses).

LO3: Discuss/present real-world implications of bias and unethical intelligent systems (Application).

C. Lab Activity

The purpose of the lab is to serve as an easily adoptable educational instrument and to support the evaluation of the project's research objectives. The lab is hosted using our

Accessible Learning Labs (ALL) [1, 32] platform, which hosts several other educational labs. To support easy classroom inclusion, the developed lab includes the components shown in Table I. The pedagogical goal of the lab is to demonstrate the impacts of inequitable software in an experiential and easily adoptable manner. This lab activity was constructed in response to the fact that ethics is becoming a more recognized need [36, 61]. Moreover, there is a lack of robust educational materials concerning the ethics of AI [51].

The lab is centered around a simple tic-tac-toe game. We chose a tic-tac-toe app since we felt that it was a game that most users would easily understand and feel comfortable using. Figure 1 shows a sample screenshot of the app.

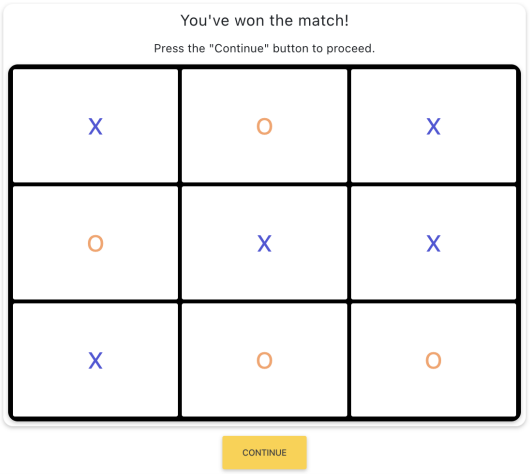

Fig. 1: Example of the tic-tac-toe game used in the study.

The user is first asked to select the avatar that most closely represents them. The user is then provided a new set of randomly generated avatars and asked to pick three avatars that they would like on their team. This is conducted under the guise that these fictitious teammates are actually other humans playing the game. This is done to bring light to the unconscious bias people have when making a selection. The activity guides the user through an avatar selection phase where they are provided with an assortment of randomly generated avatars using various characteristics (*e.g.,* face type, skin color, clothes, etc.).

Participants are then brought into a waiting room (Figure 2) where, based on their choices, an "AI matchmaking algorithm" decides to let the user into the tic-tac-toe game or gives them a penalty based on their avatars appearance or the appearance of their teammates. All participants are affected by this penalty because it is either aimed at the participant themself or a fellow teammate. This penalty (Figure 3) is in the form of a wait timer to join the game. This penalty is given to determine whether the user shows more empathy (post-survey) to those who receive a penalty based on their appearance. After the timer is up, participants are allowed to play the tic-tac-toe game where the participant interacts with a simple AI. The success or failure of their fictitious human teammates are determined randomly and then displayed to the

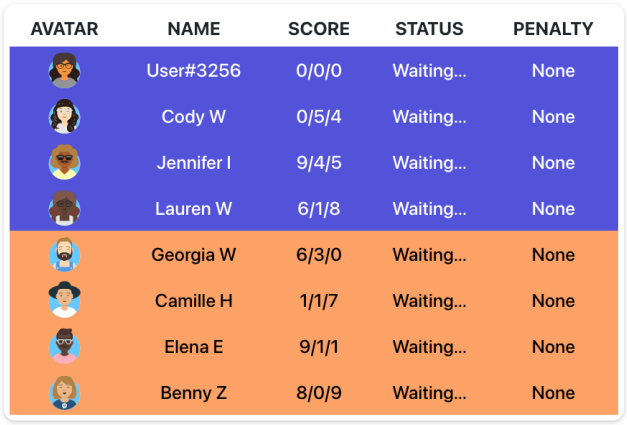

Fig. 2: Example of the match lobby used in the study. The match lobby displays the user's teammates and the team to be played against.

Fig. 3: Example of bias against the user in the form of a time delay penalty.

human participant, still under the guise that these computer-driven avatars were actually human teammates.

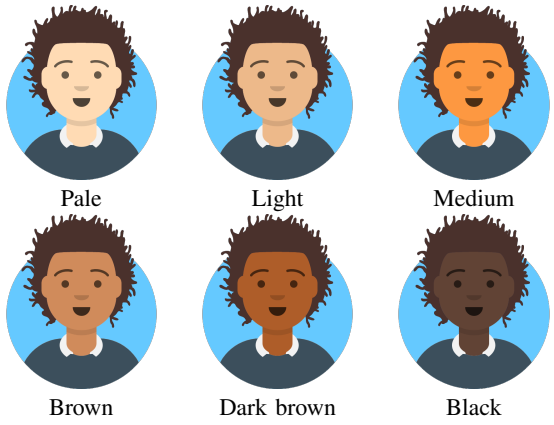

Fig. 4: Example of six different skin colors and their labels used in the lab activity.

Fictitious teammates were emulated using pre-generated avatars, designed to emulate a variety of appearances. Fictitious teammates were emulated using random-generated avatars through the alterations of ten different attributes. These attributes included hair color, skin color, clothes type, etc.. This resulted in a total of 283,295,232,000 total potential avatars that could be generated. Figure 5 displays a subset of the avatars that were used to emulate fictitious teammates in

Fig. 5: Example avatars used in application to emulate fictitious teammates.

the application. Skin colors are equally likely to be conveyed in each avatar. There is an equally likely 1/6th chance of being shown an avatar with the skin colors show in Figure 4.

No other details regarding fictitious teammates were provided to ensure that user selections were based entirely by appearance. Our analysis did not include 'men' or 'women' as avatars for several reasons. When using avatars, it is often difficult to concisely convey genders in the 'traditional' manner. Thus, there would be a large portion of participant selections that we would be unable to record the gender for. Secondly, this lab does not focus on gender identity, but rather physical identifiers. Therefore, the avatars needed to be diverse while also being generalized and androgynous.

A game match lobby before the tic-tac-toe game was used to demonstrate biases against other fictitious players. A bias against an avatar's appearance was given in certain scenarios, penalizing the user and their team with a time penalty before joining the game. This was done to create empathy for the user and demonstrate bias against others.

The self-contained nature of the lab makes it easily adoptable, with nothing that is needed to be installed or configured by the user. This will support the inclusion of this ethics-focused AI activity in a diverse set of courses, especially at resource-constrained institutions that do not have the ability to create materials on this increasingly essential topic. The experiential nature of the lab will make the activity engaging for the students, fostering interest in the topic and promoting active discussions. The lab may also be used as the foundation for bigger topics in the classroom, and provide a platform for other ethical discussions. The game was developed and evaluated using input from team members and from small informal evaluation sessions.

III. Study Design

Our study included data collected from 173 real-world participants at a local community event. Data was collected using a pre-post survey and through automated data collection mechanisms within the application.

A. Data Collection Process

Recruitment: Our human study was conducted at Imagine RIT [3], a single day event where thousands of people from the local community visit the [hidden] campus and view a variety of scientific and educational venues including robotics, software projects, and engineering activities. Visitors to the institute-wide event ranged in age 0-51+ and were generally representative of the local, non-technology-focused general population.

With the assistance of several student workers, tables were set up with laptops running the application (Section II). Participants were recruited by asking visitors passing by the tables if they would like to play a tic-tac-toe game against an AI that we had developed. Users were only provided vague details regarding the study being about AI in the event pamphlet (which very few of the participants likely reviewed or recalled information about our exhibit, since we were just one of the 100's of exhibitors at the event). Other than basic technical and process-related questions, no guidance (*e.g.,* what team members to select etc.) were provided to participants. An approved IRB was obtained prior to beginning our study.

Data Collected: Since our IRB only covered users of at least 18 years of age, we did not retain the results of anyone younger than this age that participated in our study. This resulted in a total of 173 people participating in our study who completed all required evaluation instruments. Results were collected using a pre-post survey, and an automated data collection mechanism within the lab. These actions are outlined in Figure 6 and described in the following sections.

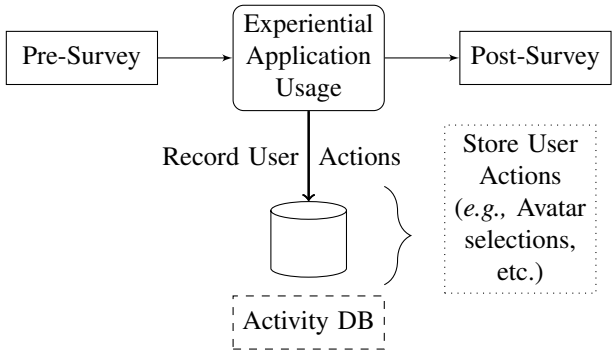

Fig. 6: Data Collection Process for In-Person User Study

B. Pre-post Survey Data Collection Instruments

Our user study comprised of three primary phases that are further described later in our work:

1) **Pre-Survey:** Collect user demographic data and rudimentary information about their AI-related feelings.
2) **Play game:** User plays the tic-tac-toe game.
3) **Post-Survey:** Users provide feedback regarding their experiences.

Participants were asked to complete pre-post surveys that are shown below. Participants were not allowed to use the application until they had completed the initial pre-survey, and they completed the post-survey only after concluding using the application. Participants who did not complete both parts

of the survey were not included in the study. The first half of the survey collected demographic information such as age and self-identified gender. The post-survey was completed after the participant had concluded the activity. This survey component measured the participant's feeling regrading bias. Due to the brevity of the large-scale in-person study, we kept the number of pre-post survey questions brief. The pre-post survey and high-level results are shown below:

Pre-Survey

1) **What Is Your Age?**
   a) 0-17 years `(n/a: excluded due to IRB)`
   b) 18-29 years `(65.9%)`
   c) 30-50 years `(12.1%)`
   d) 51+ years `(22.0%)`
2) **What Is Your Gender?**
   a) Male `(61.0%)`
   b) Female `(36.0%)`
   c) Other `(3%)`
3) **What demographic do you most closely identify with?**
   a) White `(69.9%)`
   b) Hispanic `(4.0%)`
   c) Black or African American `(2.3%)`
   d) Other `(23.8%)`

Post-Survey

1) **Were you biased against others when selecting your squad (Circle one):**
   a) Yes `(19.1%)`
   b) No `(61.3%)`
   c) Unknown `(19.7%)`

2) **How "bad" did you feel for users who are biased against. (Circle One)**

Strongly Disagree				Strongly Agree
1	**2**	**3**	**4**	**5**
1.2%	1.2%	8.7%	28.9%	60.1 %

Average response: 4.46

3) **I believe that creating unbiased software should be a top priority for software companies. (Circle One)**

Strongly Disagree				Strongly Agree
1	**2**	**3**	**4**	**5**
1.2%	1.2%	8.7%	28.9%	60.1 %

Average response: 4.46

4) **I believe that I have been impacted by inequitable/unfair software/AI in the real-world. (Circle One)**

Strongly Disagree				Strongly Agree
1	**2**	**3**	**4**	**5**
22.5%	16.2%	28.3%	17.9%	15.0 %

Average response: 2.87

C. Application Data Collection

To provide additional data metrics, participant actions were discretely automatically recorded by the application. These included actions such as the selected avatars for the fictitious 'teammates'. Each of the avatars of the fictitious teammates had demographic information associated with it (*e.g.,* skin color) that was recorded. This data was collected in a database and utilized in the analysis (Section IV-A). These database stored results were correlated with the pre-post survey data using a unique identifying value for each participant.

D. Overview of Collected Data

In addition to the collected data that has already been reported in Section III-A, we will next provide an additional breakdown of the collected data. Table II and Table III provide an overview of our collected data.

	Self-Identified Gender			
	Male	**Female**	**Other**	**Total**
White	75	42	4	121
Black/Hispanic/Other	21	30	1	52
Total	96	72	5	173

TABLE II: Participant Self-identified Gender and Ethnicity

Gender	**Q2**	**Q3**	**Q4**
M	3.12	4.39	2.81
F	3.51	4.59	2.94
Other	3.20	4.20	3.20
Ethnicity	**Q2**	**Q3**	**Q4**
White	3.17	4.46	2.64
Hispanic/Black	3.71	4.71	3.43
Others	3.46	4.40	3.38
Total	3.26	4.46	2.87

TABLE III: Average score for selected post-survey questions broken down by demographics

IV. Evaluation

Our work addresses the following research questions:

RQ1. *Did participants' demographic affect their experience with inequitable/unfair software/AI?* We observed that participants from minority ethnicity groups feel stronger regarding being impacted by inequitable software/AI.

RQ2. *Did interest in machine learning/AI affect participants' perception of the priority of unbiased software?* We observed that participants' interest in machine learning/AI positively correlated with their perception regarding the priority of unbiased software.

RQ3. *Did participants exhibit biased behavior when choosing teammates?* We observed avatars with darker skin colors were selected less frequently. Avatars with the 'black' skin color were the least likely to be selected.

RQ4. *Did the participant's demographic impact bias actions in the application?* We observed that participants from difference demographic groups exhibit similar biased behaviors.

A. Analysis Results

R1. *Did participants' demographic affect their experience with inequitable/unfair software/AI?*

To answer this research question, we examine participants' response to the post-survey question "I believe that I have been impacted by inequitable/unfair software/AI in the real-world". We are interested to see if people in specific demographic groups have stronger feeling than others regarding this question of equity.

Specifically, we are interested in whether "Gender" and "Ethnicity" can affect people's experiences and opinion of being impacted by inequitable software/AI. To this end, we utilize ordinal regression to check if "Gender" and "Ethnicity" can improve the model's fitness. We also group participants from Hispanic/Black/other into one group since most of our participants are white. Table IV shows that comparing to null model, "Ethnicity" provides significant better fitness while "Gender" doesn't provide significant improvement over "Ethnicity". Hence, "Ethnicity" is the only factor that should be included in the model specification.

Model	Test	Pr(Chi)
1 (only intercept)	NA	NA
Ethnicity	1 vs 2	0.0009456
Gender + Ethnicity	2 vs 3	0.7009207

TABLE IV: Likelihood Ratio Test for Different Specification

We used the Brant test to confirm that the parallel assumption was not violated. Therefore, we can utilize ordinal regression to model the relationship between participants' experience in inequitable software/AI and their "Ethnicity". The output of the model suggests that people from the Hispanic/Black/other ethnicities are more likely to experience inequitable software.

Additionally, we also observe a significant correlation between the participant's feeling bad about their 'teammates' being biased against (post-survey Q2) and the prevalence of the participant feeling that they had been biased against using software in the real-world (post-survey Q4). This correlation can be confirmed by Kendall's τ correlation test with p-value of 0.028: the test result suggests that there is a mild correlation between these two feelings, as the Kendall's τ equals 0.135 approximately. This may suggest that people's experience with inequitable software/AI make them more empathetic towards people being biased against.

To summarize, the primary findings of this research question include:

- Participants from minority ethnic groups had a higher proportion of prior negative experiences with software equity.
- Male and female participants share similar feeling regarding inequitable software/AI.
- There is a significant correlation between the feeling regarding inequitable software/AI and the feeling of people being biased against.

RQ2. *Did interest in machine learning/AI affect participants' perception of the priority of unbiased software?*

To answer this research question, we examine the participants' response to the post-survey question, "I believe that creating unbiased software should be a top priority for software companies." During the experiment, every participant rated to what degree they agree with the statement using a Likert range of 1 to 5.

We utilize ordinal regression to study the factors that may impact participants' perceptions of the priority of unbiased software. To this end, we use a likelihood ratio test to determine whether incorporating *Interest* (*i.e.,* participants' interest in ML/AI, obtained from pre-survey) will provide better fitness to the data. Furthermore, we expand the model specification to include *Gender* (of the participants) to investigate whether different demographics have different perception regarding priority of unbiased software.

From Table V, we see that participant identified *Gender* and *Ethnicity* is not significant compared with null model (only intercept, *i.e.,* fitting a constant model), while *Interest* provides a better fitness. Therefore, there is no significant evidence suggesting that there is a difference in demographics for perception regarding the priority of unbiased software. These results can also be confirmed by the Kruskal-Wallis test: there is no difference across demographic groups. Hence, *Interest* is the only factor that should be included in the model specification.

To further test the validity of the model, we use the Brant test to confirm that there is no violation in the parallel assumption. Therefore, it is reasonable to use ordinal regression to model participants' response against participants' interest in machine learning/AI. Lastly, the model output shows that *Interest* is positively correlated with the perception regarding the priority of unbiased software.

Model	Test	Pr(Chi)
1 (only intercept)	NA	NA
Gender	1 vs 2	0.1082466
Gender + Ethnicity	2 vs 3	0.9183158680
Gender + Ethnicity + Interest	3 vs 4	0.0002978994

TABLE V: Likelihood Ratio Test for Different Specification

To summarize, the primary findings of this research question include:

- Participants with higher interest in AI/ML tend to score higher in priority of unbiased software.
- There is no significant difference across self-identified demographic groups regarding the perception of priority of unbiased software.

R3. *Did participants exhibit biased behavior when choosing teammates?*

To answer this research question, we compared the team selection outcome with a random computer-generated simulation. If we observed a significant difference between the rates of gender and skin color selected by the participant compared to the random simulation, then it would be reasonable to assert behavior bias during the teammate selection.

During the simulation, we kept every parameter setting the same as in the real experiment: we randomly generated 16 avatar profiles, then 3 avatars were randomly selected to emulate the three avatars selected by the participant. We repeated this process for each of the participants to complete a simulated experiment. To ensure well-supported results, $10,000$ simulated experiments were executed. If the frequencies of certain traits of the avatars are less than 2.5% or more than 97.5% of the simulated experiments, then we surmise there is a bias for these traits. This range is selected to simulate a conventional 95% confident interval.

(a) Example of *Commonly* Selected Teammate

(b) Example of *Uncommonly* Selected Teammate

Fig. 7: Examples of potential popular and unpopular avatars for teammates: generated using popular and unpopular traits.

Figure 7a provides examples of a popular (commonly) selected avatar, and an uncommonly selected avatar is represented in Figure 7b. The "Black" skin color avatar attribute is significantly less selected than the simulation: only 2.49% of the simulated experiments have lower frequencies. This means that if the avatar were chosen completely randomly, it would be unlikely to have the bias as shown in the data by chance. Hence, based on the experiment, we can conclude that participants selected teammates with a "Black" skin color at a disproportionately low rate.

This selection bias against "Black" skin color can be further visualized in Figure 8: we visualize the percentage of each skin color selected in $10,000$ simulations using box plots, while using vertical lines to indicate the percentage of each skin color selected in the actual user experiment. We observe the percentage of the "Black" skin color being selected is significantly to the left, indicating there is likely a bias in the participant selection behaviors.

Additionally, we are interested in whether people are aware of those biases. We first examine participants' response to the post-survey question: "Were you biased against others when selecting your squad?". Unsurprisingly, most participants (61%) answered "No" to this question. However, compared to our random simulations, participants who answer "No" still selected avatars with the skin color "Black" less frequent than 97.73% of the simulations. Therefore, it's statistical significant that participant responses to the post-survey question do not match their actual actions. This indicates that participants either don't A) Admit their bias, or B) Recognize their bias.

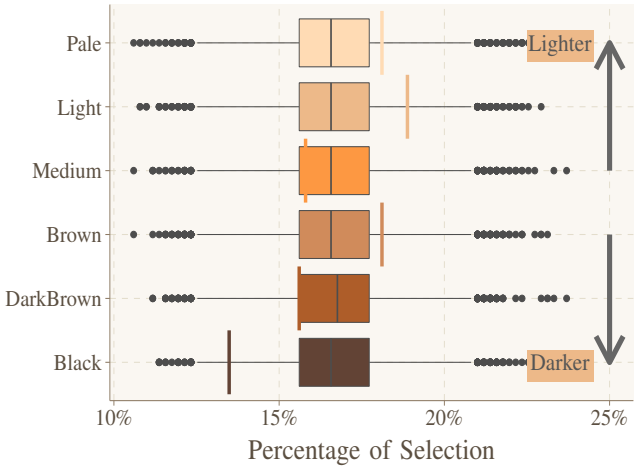

Fig. 8: Percentage of skin colors of avatars being selected during simulated experiments: vertical lines indicate the percentages in the actual user experiment. The selection behavior bias against "Black" skin color in the actual user experiment is more significant than most of random simulations.

To summarize, the primary findings of this research question include:

- Avatars with the "Black" skin color are disproportionally selected less.
- Most participants don't admit or recognize their bias.

R4. *Did the participant's demographic impact bias actions in the application?*

To answer this research question, we examine the frequency of avatars selected with completely randomized simulations, as in RQ1. We observe that male participants exhibit significant biased selection behavior for avatars with "Black" (Figure 4) skin color: there is only 2.28% of simulated experiments that have lower frequencies. Comparatively, there is no significant bias behavior in female participants. However, due to randomness, it is difficult to detect bias when data sample sizes are not large enough, meaning that we cannot definitively conclude that female participants are less biased than male participants from the above analysis. To address this issue, and further test the bias behavior across gender, we utilize the Chi-square test to see if there is a significant difference in terms of skin color. The test result shows that the p-value is 0.2627, meaning there is no significant bias across participants' gender in terms of the avatars' skin color chosen for their teammates, *i.e.,* there is no significant selecting difference between man and woman.

To further investigate the bias behavior in terms of demographics, we can further break down the participants based on ethnicity and age. Since most of our participants are white and between the age of 18-29 years old, we decided to lump the rest of the categories into other to alleviate the imbalanced data issue. We utilized the Chi-square test to check if there is a significant difference in terms of skin color during teammates selection. The test results indicate that there is no significant difference in selection behavior, meaning participants from different age groups act similarly.

In summary, the primary findings of this research question include:

- There is no significant difference in the prevalence of biased selection across different participant's demographic groups.
- Male participants exhibit significant bias against avatar with "Black" skin color during teammate selection.

V. Discussion

A. Benefits to Educators

A significant objective of computing/STEM education is to develop students' professional skills that surpass those that are frequently taught in a traditional technical curriculum [27, 56]. Ethics is a growing aspect of any computing/STEM program, and is a topic that ABET requires that all programs seeking accreditation must demonstrate [16]. Additionally, ethics is frequently viewed by students as being too ambiguous or philosophical [6]. Our created lab addresses many of these concerns by providing an easy to implement, real-world experiential ethics-focused educational component.

Publicly Available Lab Activity: There is currently a lack of robust educational materials concerning AI ethics [33, 51]. The developed experiential intervention activity will help to address this issue. This lab (Section II) is hosted on the project website [2](Figure 9) and is accessible using only a web browser. There is nothing for the adopter to install. An instructor may utilize the hosted material in a myriad of manners. For example, the activity can form the basis for a discussion regarding ethics and equitability in AI/ML. Instructors could have students use the activity to then discuss the student's feelings and observations regarding what it was like to be biased against. Such a discussion can form the basis for an ethics-focused classroom activity. Ethics and equity-focused educational material such as this is important since ethics in computing is becoming a more recognized need [36, 61]. However, there is a lack of robust educational materials concerning the ethics of AI [51].

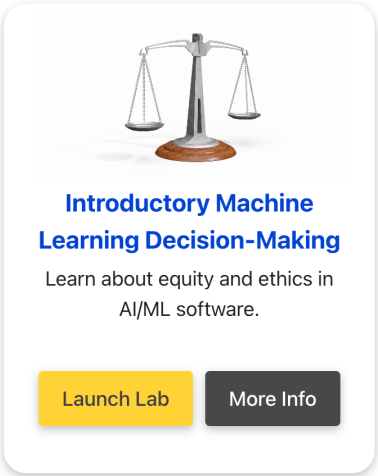

Fig. 9: Experiential Web-based Intervention Activity is Available on Project Website [2].

The activity has been used in several pilot classroom activities at both the host institution and collaborating intuitions. Preliminary participant and instructor observations indicate the pedagogical effectiveness of the intervention in an educational setting. However, pedagogical research findings regarding the inclusion of the activity in a learning environment is not yet available due to the limited participant sample size.

As AI grows in prominence, instructors are seeking ways to include this increasingly popular topic into their curriculum [54]. This self-contained activity offers an easy, resource-friendly way for instructors to include this topic into their curriculum.

B. Benefits to Individual Learners

Participants from across the world will have the capability to benefit from the lab, regardless of whether they are enrolled in a course or have instructional support. The hosted, self-contained nature of the lab will enable these individual learners to easily learn about the implications of bias in AI/ML.

C. Benefits to Software Development/Actionable Outcomes

Today's students will be tomorrow's software developers. Even a small amount of increased student awareness of bias software that can transcend into their careers as developers is likely to have a positive increase on society as a whole. The self-contained nature of the lab and brief ($\approx$ 30 minute) intervention time is expected to promote usage of the lab.

Ensuring that software developers recognize bias in both themselves and the users of software is a paramount concern for developing equitable software. Biases can make software less equitable, but also make software more difficult to use for individuals from certain groups [43].

A lack of empathy among software developers has been attributed to the creation of biased, inequitable software [11, 18]. Research has demonstrated that increasing empathy can lead to software that is developed in a more accessible, inclusive and equitable manner [1, 12, 32]. An objective of the created lab is to foster participant empathy towards groups that have been unfairly impacted by biased software.

Improved knowledge regarding empathy-creating interventions can directly benefit computing education while exponentially benefiting society through the creation of more fair, unbiased, and inclusive software used by the general population [48, 57].

Our large in-person study involving real-world users demonstrated that participants do not act in an equitable manner, even when performing a fairly mundane task with very little presumed risk or reward. This further demonstrates that developers should be cognizant of user bias and even bias within themselves [41, 43] when developing software. While there is no magical silver bullet to address user bias, some common steps to help alleviate user bias include platform design considerations (*e.g.,* matching algorithms, community policies, message and Search, Sort, and Filter Tools, etc.) [35].

D. Limitations and Future Work

There are several threats and areas of potential improvement for our work. Although participants were instructed to use the laptops in the study just like they would their own device, they were still using a laptop outside their typical environment. This means that the participant feedback and results may not properly represent what would be observed in the real world.

While we had a significant number of participants in our study (173), this still represents a very small minority of the population. The vast majority of participants were also local to the Rochester, NY metropolitan area, and may not be indicative of the entire world. An online study such as one ran on M-Turk could augment our work and provide further information. However, we believe that this would not be likely to yield any new, substantial findings. A large percentage (23.8%) of our respondents identified as an 'other' ethnicity. A ratio this large has the potential to possibly distort the findings.

Approximately 4% of the population is colorblind, and is something that predominately affects males [40]. Therefore, it is reasonable to assume a similar proportion of study participants were also colorblind. To determine the impact of being colorblind and the ability to discern the avatar's skin colors, we utilized a color blindness simulator [4]. Using this simulator, we were still able to discern essentially all avatar skin color variations with various colorblindness types (*e.g.,* Deuteranopia, Tritanomaly, Protanomaly). Therefore, we do not believe that our results were impacted by the inclusion of any colorblind participants. However, we acknowledge that no simulation tool can ever truly replicate the experiences of a colorblind individual.

During the activity, avatars were placed in a computer generated random order to limit the impacts of framing bias [7]. However, due to the inherent nature of randomness it is possible that avatars with specific demographic attributes were disproportionately placed into a position that was more likely to be selected due to framing bias [7], therefore distorting the results. However, due to the magnitude of the study (173 participants), we believe that this is unlikely to have been a problem. Additionally, this is an inherent problem with any randomly generated and located items.

We utilized avatars and not actual photographs of the fictitious teammates. This was done because we wanted to make the participant believe that they were playing with actual humans. We did not take an actual photograph of the participant and insert it into the game. Therefore, we did not feel that it would have been reasonable to assume that the participant would have an avatar while the other players had actual photos. We felt that this could lead to questions and concerns from the participant about if they were actually interacting with other actual humans.

Due to the brevity and format of the data collection activity (Section III), we did not conduct an analysis to evaluate the educational effectiveness of the created lab. Future work should be conducted to determine the lab's effectiveness for informing and motivating participants regarding bias in AI/ML. This could be accomplished in a variety of settings (*e.g.,* classrooms, outreach events, etc.) for a myriad of audiences includes computing and non-computing-focused participants, and various age ranges (9-12, undergraduate, graduate, etc.).

The lab intervention would benefit from a larger, more robust pedagogical evaluation. We have included the lab in several undergraduate and graduate courses, but a numerically significant number of results have yet to be collected to provide a sufficiently robust pedagogical evaluation. We have observed that the lab fostered student discussion regarding the topic of bias in intelligent systems. Instructors have appreciated the lab due to its simple, self-contained inclusion into their courses.

To provide more context to our results, an in-person lab study could be conducted. In this study, further participant actions and opinions could be recorded and analyzed.

Our study did not analyze gender bias, as it was not an objective of the developed lab. Additionally, androgynous characters were used due to challenges of generating traditionally gendered avatars and to not deteriorate the message that was being conveyed, since we are studying AI. Future works that wish to address this concern may perform a similar study to what we have developed, but with more traditionally gendered avatars.

Our Accessible Learning Labs (ALL) project already contains several labs that focus on software equitability. However, this is our first (and the first known in general) hosted, experiential lab that focuses on bias in AI/ML. There are a myriad of topics focusing upon equitability and inclusiveness that should have labs created regarding them. Some of these

topics include gender equitability, making ethical decisions and the benefits of a diverse multicultural development team.

Due to the number of participants and the number of possible avatars, we did not test whether certain combinations of properties will lead to bias. Moreover, most participants come from a single demographic group, *i.e.,* imbalanced data. This will likely cause bias in testing because the larger the number of each group, the more likely we can detect a significant bias. Further, there is no well-defined judging criteria that can be used to determine whether a certain participant is biased; we can only evaluate bias as a group behavior. Moreover, the non-significant results from our analysis may be due to our sample size not being large enough to identify their small effect.

Several interesting findings, such as participants with higher interest in AI/ML have a higher belief for the priority of unbiased software, in Section IV demonstrates several interesting correlations. However, we still do not fully understand the underlying mechanism for these relationships. Further work may still be required before we can apply these findings.

Our analysis was based on information extracted from single questions (Likert scale), *e.g.,* participants' feelings regarding inequitable software/AI. Consequently, the response to these questions may not be robust and faithful representations of the factors we care about. In further studies, more robust evaluation of participants aptitudes may be required to obtain more reliable data.

VI. Related Work

A. Ethical Concerns and Bias in AI/ML Education

Ethics in computing is becoming a more recognized need [36, 61]. In the United States, Computer Science programs are required to include ethics in their curriculum for accreditation. However, the manner, quality, and quantity of ethics education is left to the institutes and professors to determine [25]. Moreover, there is a lack of robust educational materials concerning the ethics of AI [33, 51].

Proponents of AI/ML ethics education contend that when students learn ethics as an integrated component in their curriculum, this important topic is taken out of isolation and is formalized [21]. There are also questions relating to the optimal manner to include ethics into the curriculum. Some educators contend that project-based learning helps students to conceptualize the real-world societal impact of bias [5]. Others contend that science fiction can be used to motivate students to consider future technologies and the ethical concerns that are likely to arise in the future [13, 25].

Fortunately, there have been several recent efforts focusing on the inclusion of ethics into AI/ML education [24, 60]. Garrett et al. [25] explored two pathways for the inclusion of ethics into AI education, including: I) standalone AI ethics courses, and (II) integrating ethics into technical AI courses. This work focused on mining syllabi on existing courses. Holmes et al. [34] discussed issues surrounding ethics and AI education by interviewing 17 AI education community members to better understand the demands and challenges of ethics in AI education. This work contends that a community-wide framework focusing on ethics in AI education using a multidisciplinary approach with a developed set of guidelines is necessary. Our work differs from these efforts in that our work is the first known work to deliver a hosted, experiential activity focusing on bias in AI/ML.

There have also been calls for AI ethics education in a variety of non-computing domains, such as for medical students [37] and even for policy strategists [53]. Aydemir and Dalpiaz [8] proposed 'ethics-aware SE', a variation of software engineering where ethical values of the stakeholders (developers and users) are collected, analyzed, and reflected upon in software specifications and in the SE processes. An analytical framework to support stakeholders in identifying ethical issues was proposed.

Experiential empathy-creating interventions have been explored in various non-computing domains, such as in medicine [23], and for creating tolerance in social situations [15]. Research demonstrates that people frequently fail to empathize with a particular target group because they are unwilling to empathize [63]. Fortunately, research suggests that empathy can be developed, often through experiential activities [1, 58]. A challenge in driving people to empathize is 'avoidance motives' [39]. An example avoidance motive is when people believe that addressing empathy-created concerns will be too costly [14] or painful [20]. Therefore, when striving to create empathy, it is imperative to demonstrate how empathy will align with, and not obstruct, the project's goals [28].

There are generally at least three related but distinct sub-processes that comprise empathy [58]. 'Mentalizing' is the ability to draw inferences about a target's feelings and thoughts. 'Experience sharing' is when a person vicariously experiences another person's emotional state [30]. 'Empathic concern' focuses on a perceiver's desire to alleviate the target's distress [9]. There are several forms of empathy, including *cognitive*, *emotional*, *affective*, and *somatic* [45]. This lab focuses on creating cognitive empathy, since it is the form that is most amiable to a computing-oriented experiential environment. An objective of the created lab is to address many of these empathy-related challenges, specifically reducing 'avoidance motives' by demonstrating the need to reduce bias in software.

There are two primary forms of empathy interventions – *Experience-based* and *Expression-based* interventions. Experience-based interventions often allow the perceiver to encounter a scenario through the target's perspective, using either a hands-on or theoretical activity. This form of intervention has been traditionally used to build empathy through a more in-depth understanding of the target's thoughts and feelings [58]. Examples of such interventions involve medical students staying in a hospital overnight to experience a hospitalization from a patient's perspective [59], or asking participants to imagine life and feelings of a member of a stigmatized group [10]. Expression-based interventions teach participants to recognize the internal states of the participant and respond appropriately. These interventions are frequently implemented in scenarios where it is difficult to identity distress in others, or when a

perceiver is impaired in conveying empathy for a target [58]. Expression-based interventions have been used in a variety of areas, such as in medical students identifying when a patient is in pain [52], and helping autistic adolescents improve their affective empathy by recognizing emotional traits in others [17].

Our ethics-focused lab utilizes experiential learning, a structure that has demonstrated its effectiveness [38]. Similar to existing efforts [1, 22, 32, 50], this lab activity enables participants to experience the addressed topic (bias in this case) firsthand. While the use of experiential learning in our work is not unique, its focus on bias is.

B. User Bias in AI/ML Software

Unconscious (aka implicit) bias has been extensively recognized and studied in a variety of areas and publications [26, 64]. Some areas in which bias has been explored include the terminology used in job postings [47], communication styles on professional networking sites [55], and even how software development is created [41, 42]. Our work did not attempt to differentiate between explicit (intentional) bias and unconscious bias. Our research only examined the rates of bias deriving from and impacting various demographics.

Peng et al. [49] studied gender bias and representation criteria. Our work did not examine gender bias, instead focusing on skin color. This previous work was conducted by having users rank potential hires. This work found that varying gender proportions can help to mitigate biases for some professions where the world distribution is skewed. Similar to our research, this work also found that the gender of the decision-maker can impact the final decision. Our work differed in that we did not examine gender bias, instead focusing on skin color.

Research has found that a lack of empathy among software developers has been attributed to the creation of biased, inequitable software [11, 18]. The created educational lab seeks to at least partially address this lack of empathy by experientially demonstrating the impact of AI/ML bias.

VII. Conclusion

This work advances both the body of knowledge regarding user actions and bias, as well as provides an experiential educational lab focusing on bias in AI-focused systems. Our key findings are that: I) Participants from minority ethnic groups have stronger feeling regarding being impacted by inequitable software/AI, II) Participants with higher interest in AI/ML have a higher belief for the priority of unbiased software, III) Users do not act in an equitable manner, as avatars with 'dark' skin color are less likely to be selected, and IV) Participants from different demographic groups exhibit similar behavior bias. The created lab and related materials are publicly available on the project website: https://all.rit.edu.

Acknowledgements

This material is based upon work supported by the United States National Science Foundation under grant #1825023, #2111152 and #2145010

References

[1] Hidden for anonymous review.

[2] Accessible learning labs. http://all.rit.edu.

[3] Imagine rit. https://www.rit.edu/imagine/.

[4] Coblis — color blindness simulator. https://www.color-blindness.com/coblis-color-blindness-simulator/. URL https://www.color-blindness.com/coblis-color-blindness-simulator/.

[5] S. Ali, B. H. Payne, R. Williams, H. W. Park, and C. Breazeal. Constructionism, ethics, and creativity: Developing primary and middle school artificial intelligence education. In *International workshop on education in artificial intelligence k-12 (eduai'19)*, pages 1–4, 2019.

[6] E. Alpay. Student-inspired activities for the teaching and learning of engineering ethics. *Science and engineering ethics*, 19(4):1455–1468, 2013.

[7] A. Aouad and D. Segev. Display optimization for vertically differentiated locations under multinomial logit preferences. *Management Science*, 67(6):3519–3550, 2021.

[8] F. B. Aydemir and F. Dalpiaz. A roadmap for ethics-aware software engineering. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 15–21. IEEE, 2018.

[9] C. D. Batson. These things called empathy: eight related but distinct phenomena. *MIT press*, 2009.

[10] C. D. Batson, M. P. Polycarpou, E. Harmon-Jones, H. J. Imhoff, E. C. Mitchener, L. L. Bednar, T. R. Klein, and L. Highberger. Empathy and attitudes: Can feeling for a member of a stigmatized group improve feelings toward the group? *Journal of personality and social psychology*, 72(1):105, 1997.

[11] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, S. Mehta, A. Mojsilovic, and S. Nagar. Think your artificial intelligence software is fair? think again. *IEEE Software*, 36(4):76–80, 2019.

[12] C. L. Bennett and D. K. Rosner. The promise of empathy: Design, disability, and knowing the" other". In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13, 2019.

[13] E. Burton, J. Goldsmith, and N. Mattei. How to teach computer ethics through science fiction. *Communications of the ACM*, 61(8):54–64, 2018.

[14] C. D. Cameron and B. K. Payne. Escaping affect: how motivated emotion regulation creates insensitivity to mass suffering. *Journal of personality and social psychology*, 100(1):1, 2011.

[15] G. L. Clore and K. M. Jeffery. Emotional role playing, attitude change, and attraction toward a disabled person. *Journal of personality and social psychology*, 23(1):105, 1972.

[16] A. E. A. Commission et al. Criteria for engineering programs: Effective for evaluation during the 2016-2017 accreditation cycle. *Website Address: www. abet. org*, 2015.

[17] M. R. Dadds, A. J. Cauchi, S. Wimalaweera, D. J. Hawes, and J. Brennan. Outcomes, moderators, and mediators of empathic-emotion recognition training for complex conduct problems in childhood. *Psychiatry research*, 199 (3):201–207, 2012.

[18] J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, Oct. 2018. URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[19] J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of Data and Analytics*, pages 296–299. Auerbach Publications, 2018.

[20] M. H. Davis, K. V. Mitchell, J. A. Hall, J. Lothert, T. Snapp, and M. Meyer. Empathy, expectations, and situational preferences: Personality influences on the decision to participate in volunteer helping behaviors. *Journal of personality*, 67(3):469–503, 1999.

[21] E. Eaton, S. Koenig, C. Schulz, F. Maurelli, J. Lee, J. Eckroth, M. Crowley, R. G. Freedman, R. E. Cardona-Rivera, T. Machado, et al. Blue sky ideas in artificial intelligence education from the eaai 2017 new and future ai educator program. *AI Matters*, 3(4):23–31, 2018.

[22] Y. N. El-Glaly, A. Peruma, D. E. Krutz, and J. S. Hawker. Apps for everyone: Mobile accessibility learning modules. *ACM Inroads*, 9(2):30–33, Apr. 2018. ISSN 2153-2184. doi: 10.1145/3182184. URL http://doi.acm.org.ezproxy.rit.edu/10.1145/3182184.

[23] J. M. Frank, L. B. Granruth, H. Girvin, and A. VanBuskirk. Bridging the gap together: Utilizing experiential pedagogy to teach poverty and empathy. *Journal of Social Work Education*, pages 1–14, 2019.

[24] H. Furey and F. Martin. Ai education matters: a modular approach to ai ethics education. *AI Matters*, 4(4):13–15, 2019.

[25] N. Garrett, N. Beard, and C. Fiesler. More than "if time allows": The role of ethics in ai education. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 272–278, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375868. URL https://doi.org/10.1145/3375627.3375868.

[26] B. Gawronski. Six lessons for a cogent science of implicit bias and its criticism. *Perspectives on Psychological Science*, 14(4):574–595, 2019.

[27] M. Ghorbani, A. A. Maciejewski, T. J. Siller, E. K. Chong, P. Omur-Ozbek, and R. A. Atadero. Incorporating ethics education into an electrical and computer engineering undergraduate program. In *2018 ASEE Annual Conference & Exposition*, 2018.

[28] A. M. Grant and D. A. Hofmann. It's not all about me: motivating hand hygiene among health care professionals by focusing on patients. *Psychological science*, 22(12): 1494–1499, 2011.

[29] M. Harris. The educational digital divide: A research synthesis of digital inequity in education. *The Effects of Brief Mindfulness Intervention on Acute Pain Experience: An Examination of Individual Difference*, 1:1689–1699, 2015.

[30] E. Hatfield, J. T. Cacioppo, and R. L. Rapson. Emotional contagion. *Current directions in psychological science*, 2(3):96–100, 1993.

[31] B. Herold. Poor students face digital divide in how teachers learn to use tech. https://www.edweek.org/ew/articles/2017/06/14/poor-students-face-digital-divide-in-teacher-technology-training.html, 2017.

[32] Hidden. Hidden.

[33] W. Holmes, M. Bialik, and C. Fadel. *Artificial intelligence in education: Promises and implications for teaching and learning*. Center for Curriculum Redesign Boston, MA, 2019.

[34] W. Holmes, K. Porayska-Pomsta, K. Holstein, E. Sutherland, T. Baker, S. B. Shum, O. C. Santos, M. T. Rodrigo, M. Cukurova, I. I. Bittencourt, et al. Ethics of ai in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, pages 1–23, 2021.

[35] J. A. Hutson, J. G. Taft, S. Barocas, and K. Levy. Debiasing desire: Addressing bias & discrimination on intimate platforms. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–18, 2018.

[36] P. Karoff. Embedding ethics in computer science curriculum, Jan. 2019. URL https://news.harvard.edu/gazette/story/2019/01/harvard-works-to-embed-ethics-in-computer-science-curriculum/.

[37] G. Katznelson and S. Gerke. The need for health ai ethics in medical school education. *Advances in Health Sciences Education*, 26(4):1447–1458, 2021.

[38] S. Krusche, A. Seitz, J. Börstler, and B. Bruegge. Interactive learning: Increasing student participation through shorter exercise cycles. In *Proceedings of the Nineteenth Australasian Computing Education Conference*, ACE '17, pages 17–26, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4823-2. doi: 10.1145/3013499.3013513. URL http://doi.acm.org.ezproxy.rit.edu/10.1145/3013499.3013513.

[39] Z. Kunda. The case for motivated reasoning. *Psychological bulletin*, 108(3):480, 1990.

[40] J. Liu. Color blindness & web design. https://www.usability.gov/get-involved/blog/2010/02/color-blindness.html, March 2010.

[41] C. Macnab and S. Doctolero. The role of unconscious bias in software project failures. In *International Conference on Software Engineering Research, Management and Applications*, pages 91–116. Springer, 2019.

[42] S. Matthiesen, P. Bjørn, and C. Trillingsgaard. Implicit bias and negative stereotyping in global software development and why it is time to move on! *Journal of*

Software: Evolution and Process, page e2435, 2022.

[43] J. McIntosh, X. Du, Z. Wu, G. Truong, Q. Ly, R. How, S. Viswanathan, and T. Kanij. Evaluating age bias in e-commerce. In *2021 IEEE/ACM 13th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE)*, pages 31–40, 2021. doi: 10.1109/CHASE52884.2021.00012.

[44] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[45] A. Mehrabian and N. Epstein. A measure of emotional empathy. *Journal of personality*, 1972.

[46] T. J. Misa. Dynamics of gender bias in computing. *Communications of the ACM*, 64(6):76–83, 2021.

[47] H. Morzogh and R. Asad. Understanding ui in crowdsourcing by mitigating unconscious bias in job advertisement.

[48] I. Niculescu, H. M. Hu, C. Gee, C. Chong, S. Dubey, and P. L. Li. Towards inclusive software engineering through a/b testing: A case-study at windows. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 180–187. IEEE, 2021.

[49] A. Peng, B. Nushi, E. Kıcıman, K. Inkpen, S. Suri, and E. Kamar. What you see is what you get? the impact of representation criteria on human bias in hiring. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 125–134, 2019.

[50] A. Peruma, S. A. Malachowsky, and D. E. Krutz. Providing an experiential cybersecurity learning experience through mobile security labs. In *International Workshop on Security Awareness from Design to Deployment*, SEAD 2018, New York, NY, USA, 2018. ACM.

[51] K. Pretz. Why schools are getting more serious about teaching engineering students about ethics. http://theinstitute.ieee.org/ieee-roundup/blogs/blog/why-schools-are-getting-more-serious-about-teaching-engineering-students-about-ethics, 2018.

[52] H. Riess, J. M. Kelley, R. W. Bailey, E. J. Dunn, and M. Phillips. Empathy training for resident physicians: a randomized controlled trial of a neuroscience-informed curriculum. *Journal of general internal medicine*, 27(10): 1280–1286, 2012.

[53] D. Schiff. Education for ai, not ai for education: The role of education and ethics in national ai policy strategies. *International Journal of Artificial Intelligence in Education*, pages 1–37, 2021.

[54] A. A. Shaikh, A. Kumar, K. Jani, S. Mitra, D. A. García-Tadeo, and A. Devarajan. The role of machine learning and artificial intelligence for making a digital classroom and its sustainable impact on education during covid-19. *Materials Today: Proceedings*, 56:3211–3215, 2022.

[55] A. Shevlin. The impact of unconscious gender bias on online professional networking & recruitment. 2018.

[56] T. J. Siller, A. Rosales, J. Haines, and A. Benally. Development of undergraduate students' professional skills. *Journal of Professional Issues in Engineering Education and Practice*, 135(3):102–108, 2009.

[57] M. Vorvoreanu, L. Zhang, Y.-H. Huang, C. Hilderbrand, Z. Steine-Hanson, and M. Burnett. From gender biases to gender-inclusive design: An empirical investigation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[58] E. Weisz and J. Zaki. Empathy building interventions: A review of existing work and suggestions for future directions. *The Oxford handbook of compassion science*, pages 205–217, 2017.

[59] M. Wilkes, E. Milgrom, and J. R. Hoffman. Towards more empathetic medical students: a medical student hospitalization experience. *Medical education*, 36(6): 528–533, 2002.

[60] R. Williams. How to train your robot: Project-based ai and ethics education for middle school classrooms. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pages 1382–1382, 2021.

[61] S. Wykstra. Fixing tech's ethics problem starts in the classroom, Feb. 2019. URL https://www.thenation.com/article/teaching-technology-ethics-big-data-algorithms-artificial-intelligence/.

[62] S. Yan, H.-T. Kao, K. Lerman, S. Narayanan, and E. Ferrara. Mitigating the bias of heterogeneous human behavior in affective computing. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

[63] J. Zaki and M. Cikara. Addressing empathic failures. *Current Directions in Psychological Science*, 24(6):471–476, 2015.

[64] N. Zelevansky. The big business of unconscious bias. *The New York Times*, 2019.