Spring 2023

UNTRAINING GENDER BIAS: AN EYE-TRACKING STUDY

Ripujit S. Bamrah

San Jose State University

UNTRAINING GENDER BIAS: AN EYE-TRACKING STUDY

A Master's Project

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfillment

Of the Requirements for the Degree

Masters of Science in Computer Science

By

Ripujit S. Bamrah

May 2023

The Designated Thesis Committee Approves the Thesis Titled

UNTRAINING GENDER BIAS: AN EYE-TRACKING STUDY

By

Ripujit S. Bamrah

APPROVED FOR THE DEPARTMENT OF COMPUTER SCIENCE

SAN JOSÉ STATE UNIVERSITY

May 2023

Nada Attar, Ph.D.	Department of Computer Science
Reem Albaghli, Ph.D.	Department of Computer Science
Thomas Austin, Ph.D.	Department of Computer Science

ABSTRACT

In recent years, social cognitive theory has emphasized the role of cognitive processes in shaping perceptions and behavior related to gender bias. By examining the impact of targeted training interventions, this study seeks to better understand the influence of such processes on decision-making in the context of character selection. This human-computer interaction study explores the potential of intervention-based training to untraining gender bias in character selection. With an increasing need to address gender bias in various domains, understanding the impact of gender-based training becomes crucial. According to our hypothesis, exposure to masculine characters would boost people's preference for female-intellectualized characters. Utilizing a two-part experiment, subjects were presented with a series of images across three blocks, with the second block providing gender-specific training. Experiment 1 focused on training with female characters, while Experiment 2 used male characters. The results demonstrated a significant increase in female character choices in Block 3 compared to Block 1, particularly for female-intellectualized characters in Experiment 2. Eye-tracking data further revealed slower response times and greater pupil size for female characters in Block 3 compared to Block 1 in both experiments, indicating higher cognitive load. These findings suggest that intervention-based training can effectively counter gender bias in character selection.

Keywords: gender bias, untraining bias, eye-tracking, male and female, cognitive load, human-computer interaction

ACKNOWLEDGMENTS

Thank you to:

Professor Nada Attar, for providing important advice and assistance during this research process. Your guidance and expertise in the field of human-computer interaction has been extremely helpful to the outcome of this study.

Mama, Papa, Supriya, and Nicole for your unwavering support and encouragement not just throughout my academic career, but in all areas of my life. Your love, guidance, and sacrifice have been essential to my growth and success as an individual.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I. INTRODUCTION

Gender biases have historically permeated all facets of society, influencing the expectation of men and women, often leading to unequal opportunities between the sexes. This often leads to unequal opportunities and limited beliefs for females. The history of this gender gap trace back to the age of agriculture when men appraised themselves for doing work while devaluing female labor [1]. More recent than that, research has shown that due to gender bias, female philosophers have turned in less articles and journals, participated in less discussions, and specialized in less research topics than the average male philosopher [2]. Given the deep-rooted nature of this bias in human history, numerous efforts have been made to address and mitigate it. With the advent of new technology, such as eye-tracking, we now have the potential to gain a better understanding of the issue and develop more effective strategies to counteract it.

Gender bias has become so rooted in society that the negative impacts of gender bias are shown constantly in many fields such as employment, education, politics, and even personal relationships. Studies suggest that gender bias can be exhibited by individuals of all genders, and researchers are actively working to better understand and address this issue. It is vital to find and put into practice effective ways that question and mitigate these stereotypes to reduce gender bias. Addressing these stereotypes can pave the way for more equal and diverse representations of both genders.

Eye-tracking technology can be a powerful tool in human-computer interaction research by allowing us to directly measure where an individual directs their visual attention on a screen and thereby see where their cognitive load is greatest. Underlying data about the location and duration of an individual’s eye movement, a greater understanding about cognitive processes underlying character selection will show us how attention and decision-making are influenced by intervention-based training. This field has provided valuable insights into cognitive processes and attentional patterns that show how humans process information and make decisions by examining the relationship between visual attention, cognitive load, and decision-making behavior.

Untraining gender bias, or the process of removing biases and stereotypes, has become a focus of recent research to support more fair gender representations and mitigate harmful biases. There have been attempts to untrain gender bias, however, individuals still suffer unconscious gender bias that continues to shape their perspectives and attitudes.

Most recently, a paper by Albaghli and colleagues found that using art as a stimulus can be effective in untraining gender bias. In their study, they found a bias towards non-intellectual images for females which was diminished after their training block. This study provides insight into measuring biases. The purpose of my research is to replicate Albaghli and colleagues' findings.

In this experiment, the focus of our research was to determine whether it is possible to untrain gender bias with additional provided by eye tracking so we can see how visual attention and decision-making are influenced during the experiment. These biases continue to exist, even if an individual is making a conscious effort to reject prejudice and discrimination. In this study, we conducted two experiments: the first was to replicate Albaghli and colleagues' study to see the increase in female character choices occurs after a gender-based training intervention. We replicated this study with the primary goal of incorporating eye-tracking technology to enhance our understanding of the underlying cognitive processes. The second experiment aimed to see if the training on male characters leads to a preference for female-intellectualized characters.

Researching the effectiveness of untraining interventions in modifying gender bias also gives us knowledge on how to mitigate these biases and develop a more effective strategy for promoting diversity. An untraining intervention could be used in the workplace to reduce the impact of gender biases on decision-making processes such as hiring, promotions, and workplace culture. Having both male and female employees engage in these interventions could lead to better outcomes for women in terms of equal representation.

1.1 Hypothesis

In the first experiment, we replicated Albaghli and colleagues' gender-based training experiment to test the effect of seeing only female images during the training block on participants' choices in subsequent image selection tasks. The hypothesis in Albaghli and colleague's paper states "that participants choose a greater number of intellectualized female characters after the training block, comparing to the first block that is before the training block" [3]. The researchers used this middle training block to test whether it was possible to untrain gender bias. To replicate this, I created an experiment in MATLAB that consisted of 3 blocks. Each block consisted of 10 trials each. The second block in this experiment is meant to train the participant. Block 1 gives the participant a choice between male and female characters, Block 2 gives the user the choice of two female characters, and Block 3 will give the choice of male and female characters again.

In the second experiment, I created another project in MATLAB that consisted of 3 blocks in similar fashion to Experiment 1. Block 1 will give the user the choice of male and female characters, Block 2 will give the user the choice of two male characters, and Block 3 will give the option of two female character (female-intellectualized and female-only). We hypothesize that the participants will choose a greater number of female-intellectualized characters than female-only characters after the training block due to the preference for female-intellectualized characters will show that the participants are less inclined to connect females with attributes focusing primarily on looks. We tested the relationship between image selection and eye-tracking data by using eye-tracking technology to measure pupil size and reaction time as indicators of cognitive load and decision-making processes. We wanted to see whether subjects would be more likely to pick a female image as opposed to a male image after exposure to a training block that had two female options.

Alternative Hypothesis (H1): There will be a significant increase in the proportion of female-intellectualized choices compared to female-only choices in Experiment 2, Block 3 after the training intervention.

II. RELATED WORK

2.1 Identifying Biases

One recent popular research topic is detecting gender stereotypes in different industries and fields. Cryan et al. did an experiment on natural language and gender stereotype detection [4]. The researchers used machine learning to quantify and detect gender stereotypes. Andrich and Emise also performed an experiment on gender stereotypes and politics [5]. In this experiment, the researchers evaluated how social media affects the way an individual views politicians - associating female politicians with leadership and competence but associating male politicians with more political career-relevant traits. In this particular experiment, they analyzed Facebook comments about Donald Trump vs. Hillary Clinton in which Trump was seen as a more qualified candidate due to more masculine traits.

Another experiment was done by Matthews et al. detecting gender stereotypes when it came to legal opinion word embeddings [6]. Matthews et al. performed an experiment using machine learning and used a Natural Language Processing model to train on legal opinion texts in U.S. case law. This model recreated these biases which revealed that in these word embeddings are biases and stereotypes.

As previous studies have shown, gender bias is an issue in many areas of society - this includes the technology sector. Previous studies have shown that gender bias can arise from underrepresentation of women in the workforce to the presence of biased algorithms and technologies. The topic of gender bias has been acknowledged when it comes to the context of programming languages - as well as evidence that may indicate that they may contribute to gender bias. This section reviews research on gender bias in tech that may be contributing to the perpetuation of gender stereotypes.

A study done by Brooke investigated gender bias on Stack Overflow - an online forum used to ask questions and solve issues around programming - using 11 years of activity. Brooke used a gender identification method to examine if there were key user metrics of success (reputation points, user tenure, level of activity). The results show that females that responded received lower scores for their answers (even though they showed more effort) [7]. Feminine users interact more with other feminine users. Even in this study, it is important to note that even gender shapes interaction in technical spaces.

Zyte also did research on gender inequality across programming languages [8]. They used their own data set, and the gender of a profile was given by the company based on their own data. Zyte used two methods and analyzed programming languages (such as Python, Ruby, Java, C#, C++, JavaScript, and PHP). Percentages by language showed that the largest female percentage was around 20% when it came to Ruby.

BBC released an article about a research study that showed that a pull request made by a female had higher approval scores than those from males - only if the female's gender was unidentifiable [9]. However, if their gender is identifiable, the acceptance rate gets lower. These same researchers tried to consider whether factors such as quicker response times, or easier to comprehend played in when it came to higher acceptance rates, but there was no correlation found.

In a dissertation, Liu did research on detecting and mitigating bias in natural languages [10]. Natural Language Processing (NLP) is an area of artificial intelligence in which humans and machines can interact with one another using natural languages. NLP allows machines to comprehend and analyze natural language. Liu's dissertation concludes that NLP shows racial and gender bias but there are ways to mitigate and reduce bias to improve the fairness of NLP systems.

2.2 Mitigating biases

Gender bias and stereotypes have been persistent issues in many modern-day industries and fields. In recent years, there has been research done to try and address and mitigate these issues using technology. In one study done by Moss-Racusin et al., researchers found that gender bias contributes to the underrepresentation of women in STEM [11]. To reduce this bias, these researchers created VIDS (Video Interventions for Diversity in STEM) which are a set of videos that are aimed to increase awareness of gender bias in this area. After showing these videos to participants, there were experiments conducted to test the effectiveness of these videos. These experiments proved successful by increasing positive attitudes towards females in STEM.

In another experiment, two researchers, Kirtane and Anand, did research on natural language processing systems on two Indian languages - Hindi and Marathi [12]. In languages such as Hindi, Marathi, and even latin-root languages, there are masculine and feminine words or parts of words. Bias is measured by using a dataset that consists of neutral words paired with Embedding Coherence Testing and Relative Norm Distancing. Using these two debiasing techniques, it is possible to mitigate gender bias in both Hindi and Marathi.

Beltran et al. also ran an experiment to mitigate implicit gender bias [13]. In their research, a first-person avatar was used in a virtual environment. Scenarios would play out to the subject in the first-person avatar. The study showed that when it came to female scenarios, subjects showed a lower level of implicit gender bias. While the implementation of technology is not a complete solution for gender bias, it has the potential to be a valuable tool in the ongoing effort to promote gender equality in various fields.

2.3 Eye Tracking

As research continues in our modern day and age, many different sectors have made use of eye-tracking as a powerful tool to do extensive research. The presence of eye tracking devices has allowed us to research the effects of gender bias in multiple settings. Eye tracking allows researchers to see the exact measurement of where people look and how long. With eye tracking, researchers can better understand how gender bias appears and analyze data to address it.

Valtakari et al. developed an interest in gaze behavior when an individual is interacting with something or someone [14]. These researchers found that there are many ways that eye tracking can be done through many different interactions. Throughout the research done, there are many different setups that can be useful for different types of experiments from single eye setups to dual eye setups and freedom of movement levels. Kerr-Gaffney et al. performed an eye-tracking study on a population that has been diagnosed with eating disorders (ED). In the experiment, the researchers used eye tracking to study the biases towards different foods and body stimuli in eating disorders. The researchers then compared these studies with those who had other psychiatric disorders [15].

Another study was done by Pfiffelmann et al. in which there was examination on the effects of personalizing a job advertisement with the receiver's name and photograph [16]. The study used eye tracking to take a measurement of visual attention and through the study it was found that these participants focused harder and longer on the personalized advertisements versus those job advertisements which were not personalized.

Bogomolova et al. performed an experiment on unit pricing on products in grocery stores [2]. Unit pricing is inconsistent when it comes to information. These researchers performed a study to see how different design factors of unit price labels (such as position, font size, signposting, and color highlighting) affect a consumer's eye movements while grocery shopping as well as how the effects vary based on how conscious a consumer is of the price. Eye tracking tools revealed that an enhanced design for a unit price label led to increased eye fixations. Bogomolova et al. concluded that improving the design label for unit pricing likely has minimal effects on a user. Although there are still issues to be resolved in eye tracking research, the results to date indicate that this technology has enormous potential to improve our knowledge of human behavior and cognition.

III. EYE-TRACKING

3.1 History and Background

Eye trackers are essential tools in modern-day studies of cognitive load and human behavior, particularly within the realm of Human Computer Interaction (HCI). By utilizing eye trackers, researchers can track and analyze individuals' subjects' gaze patterns, fixation durations, and pupil size. In this section, we explore the history, development, and applications of eye tracking technology.

Early eye tracking placed participants in front of a mirror while an examiner would record their eye movements [17]. Research using this early eye tracking technique revealed that reading was nonlinear – this meant that eyes do not move continuously with the text but instead they move and jump around just staying on a word long enough to comprehend it [14].

In 1901, R. Dodge and T.S. Cline developed a more precise eye-tracking system, the Dodge Photochronograph, which used light to reflect off an individual's eyes and onto a photosensitive plate [17]. Figure 1 illustrates this early eye tracker.

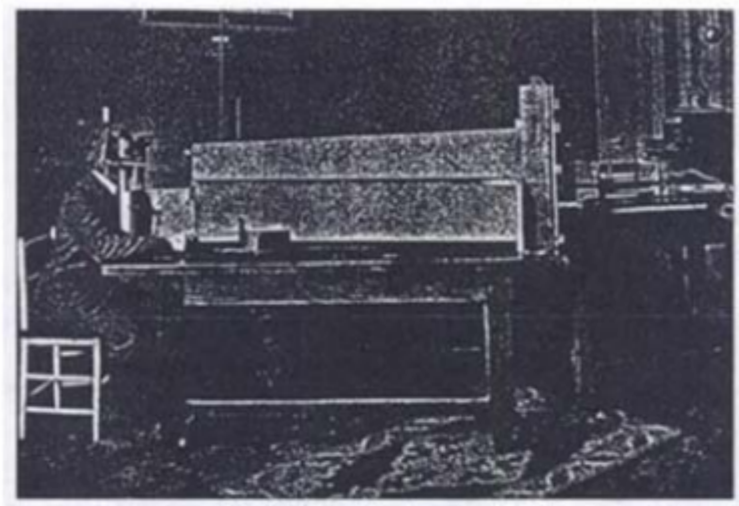

Figure 1: R. Dodge and T.S. Cline's first precise eye tracker [17]

3.2 How it Works

Eye tracking devices use near-infrared illumination to reflect light off an individual's eyes. Those reflections are then detected and analyzed by the built-in cameras in the eye tracker. Eye-trackers can be used to determine the X, Y, and Z coordinates of an individual's gaze, as well their pupil dilation, and reaction time. The device is also able to track gaze origin, gaze point, and pupil dilation for both eyes as well as calculates an origin point based on where the individual is looking on the computer screen [18].

3.3 Cognitive Load

Krejtz et al. performed an experiment in which the researchers used two different metrics to test the individual's sense of difficulty of a task – change in pupil diameter as well as the rate and magnitude of microsaccades (small involuntary eye movements during visual fixation) [19]. In the experiment, participants performed mental arithmetic ranging between easy problems to challenging problems, while focusing on a single target. Results showed that both measures can accurately indicate the level of mental effort required for a task.

The relationship between the eye-tracking measures and cognitive processing can be better understood by examining the underlying assumptions. While eye tracking does not directly measure cognitive processes, it serves as valuable information, providing insight into the extent and nature of cognitive effort. By establishing this connection, eye-tracking data can be used to make inferences about the cognitive processes taking place during decision-making, problem-solving, and other complex tasks.

Cognitive Load Theory (CLT) is how humans process information and the amount of working memory resources that are used [19]. Cognitive load is the amount of mental effort and resources needed by someone to do a specific task. This allocation of attention and mental effort can be analyzed through eye tracking because it is reflected in eye movement patterns and fixations – long fixations and rapid eye movement are just a few signs of increased processing demands. Understanding cognitive load is crucial because it can be used to enhance designs and increase learning results.

According to Cognitive Load Theory (CLT), overloading the working memory (which is limited) results in less learning [16]. The 1998 model divided the working memory into three different types: intrinsic cognitive load (ICL), extraneous cognitive load (ECL), and germane load (GCL). Intrinsic cognitive load is how difficult a topic is, regardless of the explanation of that topic. For example, whether or not a Professor explains clearly to you how to traverse a binary tree, the difficulty level of that task will remain the same - cannot be controlled. Extraneous cognitive load is the load experienced by working memory while trying to solve a task with materials. This can be somewhat controlled. Lastly, germane load is the effort required, mentally, to process and learn new information in a way that can be stored in memory and used later.

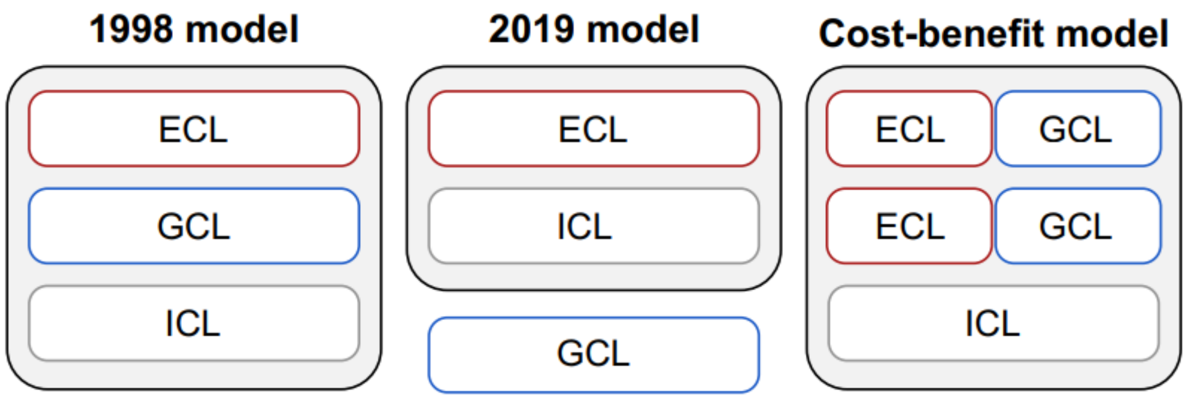

Figure 2: Comparison of revised and advanced cognitive load models [7]

Figure 2 shows a comparison of 3 models: 1998 model, 2019 model, and a cost-benefit model. The difference between the 1998 model and 2019 model is that the Germane model is thought of as load at this point, but instead as processing rather than loading [20]. The cost-benefit model introduces the idea that if a task is being performed or learned, the extraneous cognitive load and Germane load (or processing) can vary.

In an HCI study done by Sevcenko et al, the researchers had 42 participants play a time-critical game so they could measure cognitive load through eye tracking. The game had different levels of difficulty which allowed researchers to see and measure cognitive load through different gaze data points such as eye fixation frequency, saccadic rate, and pupil diameter. All these factors gave researchers the ability to predict the difficulty of a task. Through this experiment, Sevcenko et al. was able to pave the way for measuring cognitive load in different and similar situations [21].

3.3 Eye Tracking and Bias

Eye tracking has been used in research as a powerful tool to study gender and racial bias. Using eye tracking to capture eye movements and other metrics allows researchers to gain insight into attitudes and behaviors which often are difficult to capture otherwise. Recently, eye tracking has been used to reveal how these biases can be shown in gaze behavior. In a study done by Man and Hills, the researchers found gender bias in facial recognition in females only [22]. Different patterns in eye movement were found when scanning their own faces versus scanning a face of the opposite gender. Researchers found that females tend to have stronger eye movements and gaze behaviors when looking at the chin, lips, nose and eyes of another female as shown in Figure 3.

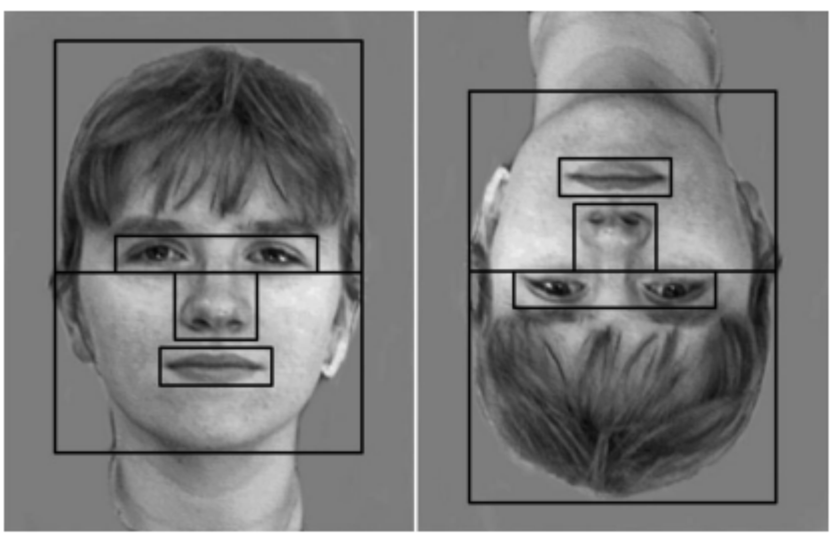

Figure 3: Areas of interest for faces [22]

In another study, Kawakami et al. performed 4 different experiments that examined the effects of intergroup motivations to the eyes of both ingroup and outgroup members [23]. By performing eye tracking, the researchers were able to find that white participants would focus on eyes when seeing other white subjects as compared to black subjects. Figure 4 shows that participants showed similar areas of interest that were captured by the eye tracker. People who belong to a certain racial group look closer and deeper at the eyes of their own group, meaning there are more biases towards their own group. Both the study on gender bias and racial bias are used to demonstrate how eye tracking technology can be used to examine them. The findings from both experiments show how people interact with people from different and like groups.

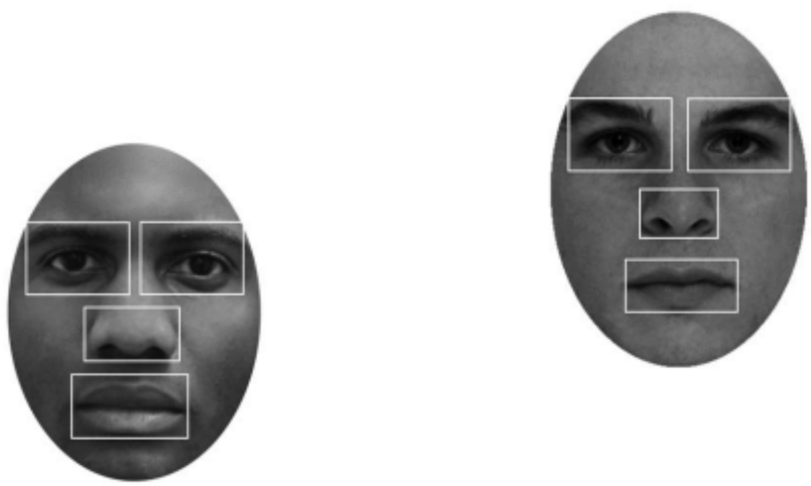

Figure 4: Areas of interest for faces [23]

The findings in these studies highlight the importance and prominence of bias in one's everyday life, and how eye tracking technology can be utilized to observe these biases. Results show that these biases can impact how individuals interact with other individuals from different social groups. The insights we gain from these studies give a better understanding about biases in different contexts and are important to know about so we can mitigate them.

IV. METHODOLOGY

4.1 Participants

A total of 16 participants (M_{age} = 23.93, *SD* = 3.18) were recruited from San Jose State University who took part in the study. Of the 16 participants, 8 were male (M_{age} = 25.5, *SD* = 3.026) and 8 were female (M_{age} = 22.5, *SD* = 3.343). All participants reported normal or corrected-to-normal vision and no medical conditions that would impair eye movement. Each subject provided written consent and received a $10 honorarium for their time. Before participating in the study, the participants were informed about the study's purpose after conducting the experiment to ensure the title and reason for experimentation would not skew the results. Participants were informed that they could withdraw from the study at any time without penalty.

4.2 Dataset

As noted by Albaghli et al., there have been few studies examining the perception of Middle Easter culture through art visuals [3]. In this study, the dataset consists of 19th-century paintings from the Orientalism era [24, 25]. Two out of three of these artists visited the Middle East. This experiment continues Albaghli and colleagues' work. Each painting was collected and analyzed based on the time, location, and rationale behind its creation. Paintings were categorized based on whether they were created by an artist who had visited the Middle East and accurately depicted cultural conventions or by an artist who had never visited the region and misrepresented these conventions.

Figure 5: Example of three characters used in the experiment (FI, MI, FO respectively)

Table 1 below describes the three types of characters used in the study as defined by Albaghli et al. [3]. Figure 5 shows an example of the characters used while Figure 6 presents a sample background image used in the experiments. The three characters are categorized as female-intellectualized, male-intellectualized, and female-only respectively.

Label	**Code**	**Description**
Female-Intellectualized	FI	A female character that is depicted performing an intellectual activity or posing in a non-sexualized manner.
Male-Intellectualized	MI	A male character that is depicted in either a dignified state or would be performing an intellectual activity.
Female-Only	FO	A female character who is presented in a stereotypical scene (such as unclothed or being sold as a slave).

Table 1: Description of Characters used in the experiment [3]

Figure 6: Example of Background Image used in the experiment

4.3 Equipment

The equipment used in this experiment consists of the Tobii Pro Fusion Eye Tracker and two computers. The Tobii Pro Fusion, shown in Figure 7 is a dual eye tracker and has a sampling frequency of 250 Hz with a 250ms gaze recovery time. The data sample output given by the eye tracker software is the system time stamp, gaze origin information, gaze point information, pupil diameter, and validity code. The eye tracker sits where the laptop computer hinges and has a length of 374 millimeters, a height of 18 millimeters, and a width of 13.7 millimeters [26]. In this experiment, we used the setup as seen in Figure 8b – free head movement.

Figure 7: Tobii Pro Fusion [26]

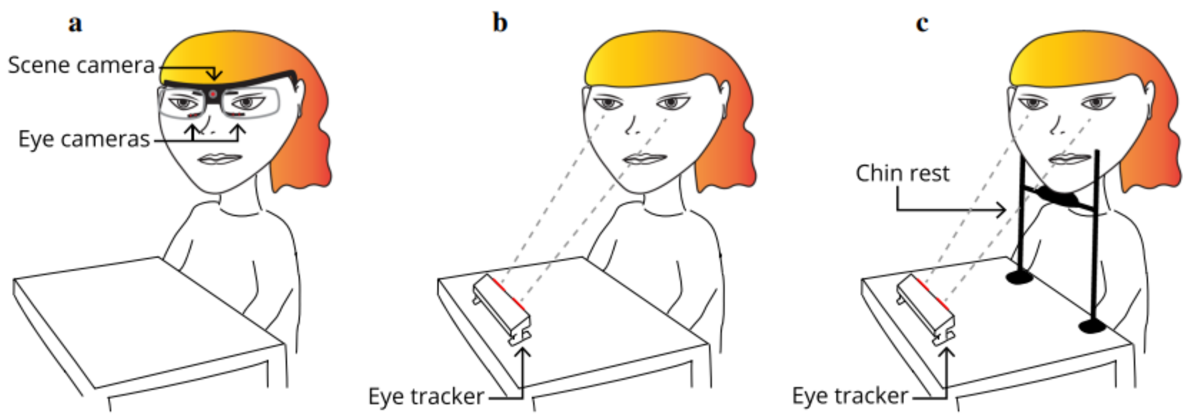

Figure 8: Different types of Eye Tracker Setups [27]

The computer that was used to conduct the experiment with the eye tracker was a 2.6 GHz Dell Latitude 7420 that has an 11th gen Intel Core i5 Processor with 4 cores, 8 threads, and 16 GB of RAM. The computer has a 14 inch screen with 1920 x 1080 resolution.

The computer that was used to create and modify the experiment script in MATLAB, analyze scripts and data in python and create visuals for the report was done on an Early 2013 MacBook Pro that has a 2.8GHz Intel Quad-Core i7 processor and 16 GB of RAM.

4.4 Experiment / Procedure

4.4.1 General Experiment Design

The experiment was designed in two parts to determine whether untraining gender bias is possible. Each part consists of three blocks, with each block having 10 trials. The **dependent variable** is eye tracking measures such as pupil size and reaction time, while the **independent variable** is the training intervention (gender-based training in Experiment 1 and male character training in Experiment 2).

4.4.2 Procedure for Each Block

For Experiment 1, participants were invited to take part in a study about gender bias. Upon arrival at the laboratory, participants were asked questions about their vision. All participants then sat at a desk facing a wall to minimize distractions with a computer in between. The eye tracker was placed on the laptop at the hinge. Participants were then calibrated with the eye tracking machine. Participants were told that there would be 3 images on screen – a character, a background image, and then another character. They were asked to choose a character based on preference. They were then given a demonstration on how to answer questions using the Z and C keys on the keyboard for the left character and right character respectively. In between each trial, there is a small cross at the center of the screen for 1 second to recalibrate the eyes. After the demonstration, participants ran a practice round to ensure understanding of the commands given. After the demonstration, the official experiment was run with the first 10 trials (Block 1) asking the character to choose between male and female characters. Block 2 asked the participant to choose between 2 female characters. Block 3 asked the participant to choose between male and female characters to observe whether there was a difference between the number of female responses in Block 1 versus Block 3. After completing

Experiment 1, the user was given a two-minute break, and then asked to do Experiment 2. Experiment 2 followed the same procedure as Experiment 1, but the choices in Block 2 were between 2 male characters, and the choices in Block 3 were between a female-intellectualized character and a female-only character. Block 1 remained the same. Finally, all participants completed the experiment. They were then debriefed about the true purpose of the study and were given a $10 honorarium for their time.

4.4.3 Eye Tracking Setup

The experiments were conducted using an eye tracker. Eye Tracking was used to measure the gaze points, gaze dilation, gaze origin, and system time stamp for each trial. The experimentation code was written in MATLAB and was wrapped in a for-loop that displayed the background and character images on screen using textures. Messages were logged between each trial to help split up the data and determine which trial each piece of data belonged to. A plus sign appeared between each trial to help re-calibrate the participant's eyes to the center of the screen.

V. ANALYSIS AND RESULTS

5.1 Overview

In this section, we present a detailed analysis of the data collected from our two-part experiment. We focus on the effects of the training interventions on character selection, response times, and pupil size. Our independent variable is the training intervention (gender-based training in Experiment 1 and male character training in Experiment 2), and our dependent variables are the eye-tracking measures, which include response numbers, response times, and pupil size.

The analyses were performed using descriptive statistics, including calculating the mean and standard deviation for each measure to provide an overview of the central tendency and variability of the data. By comparing the averages for response time, response numbers, and pupil dilation, we aim to understand how our training interventions influenced the participants' decision-making processes, and whether these changes support our hypotheses. We visualize the results using graphs to highlight the difference observed and facilitate the interpretations of our findings.

5.2 Average Female Choice Response Numbers in Exp. 1

The difference in the average number of female choices between Block 1 and Block 3 in Experiment 1 was not statistically significant for all participants ($t(30) = -0.68$, $p = 0.5$). As shown in Figure 9, the average number of female characters chosen in Block 1 was 5.25 and that number slightly increased to 5.63 in Block 3. This result indicates that there was no substantial change between Block 1 and Block 3 in choosing female characters after the training phase in Block 2.

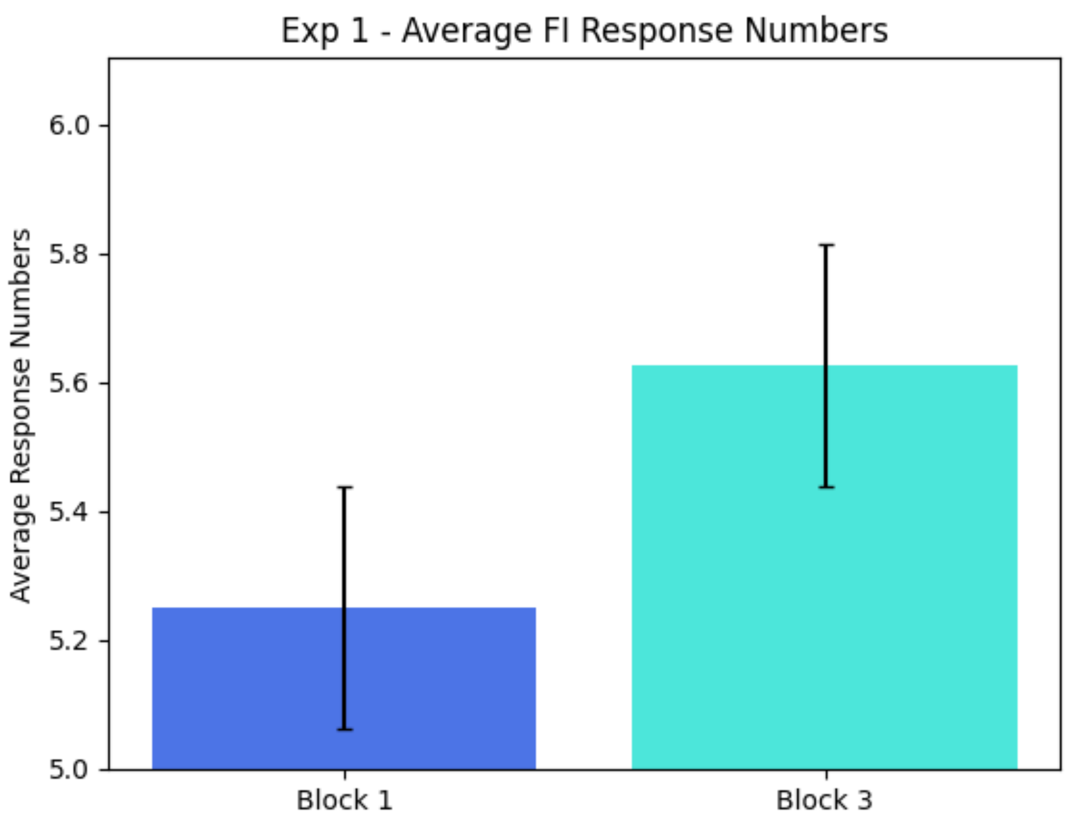

Figure 9: Average Number of FI Responses in Exp 1

5.3 Average Female-Intellectualized Choice Response Numbers in Exp. 2

The difference in the average number of female-intellectualized choices between Block 1 and Block 3 in Experiment 2 was **statistically significant** ($t(30) = -10.54$, $p < 0.0005$). As shown in Figure 10, the average number of female-intellectualized characters chosen in Block 1 was 3.0 which increased to 7.69 in Block 3 (increase of 156.33%). This result demonstrates a significant change between Block 1 and Block 3 in choosing female-intellectualized characters after the training phase in Block 2.

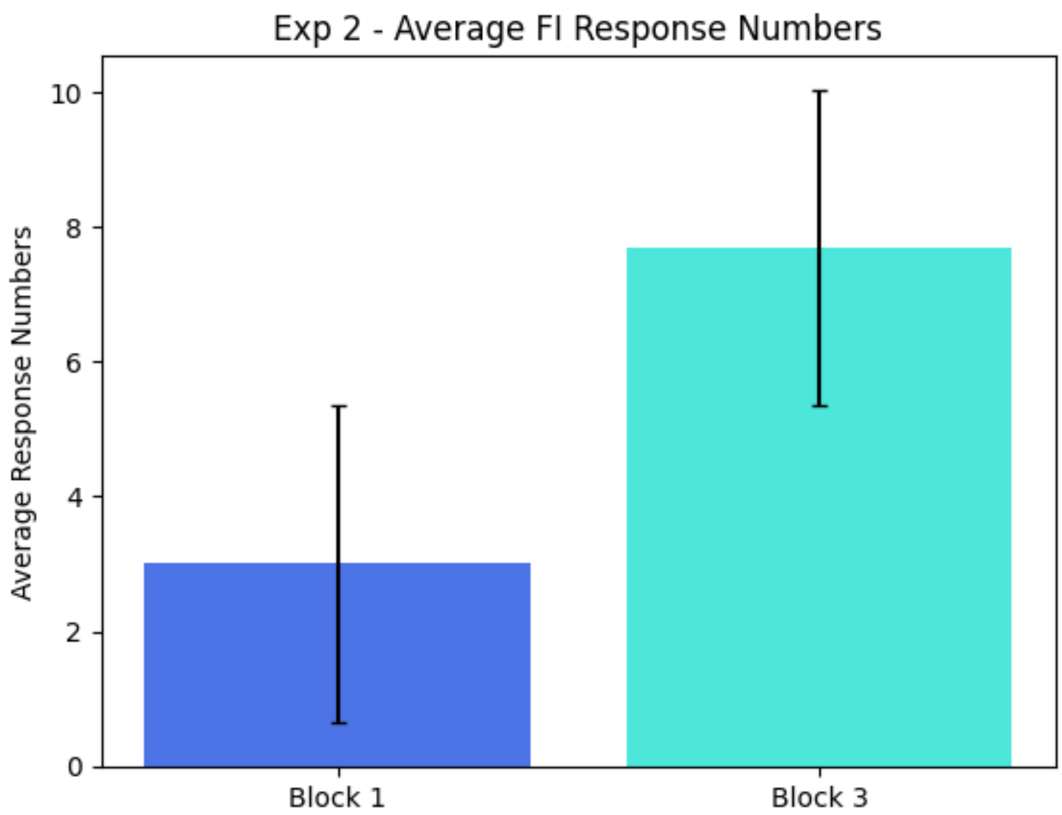

Figure 10: Average Number of FI Responses in Exp 2

5.4 Average Female Choice Response Time in Exp. 1

The difference in the average female choice response time between Block 1 and Block 3 in Experiment 1 was not statistically significant for all participants ($t(30) = -0.34$, $p = 0.74$). As shown in Figure 11, the average response time for choosing female characters in Block 1 was 4.64 seconds, which increased slightly to 4.90 seconds in Block 3. This result suggests that there was no meaningful change in response time between Block 1 and Block 3 for choosing female characters after the training phase in Block 2.

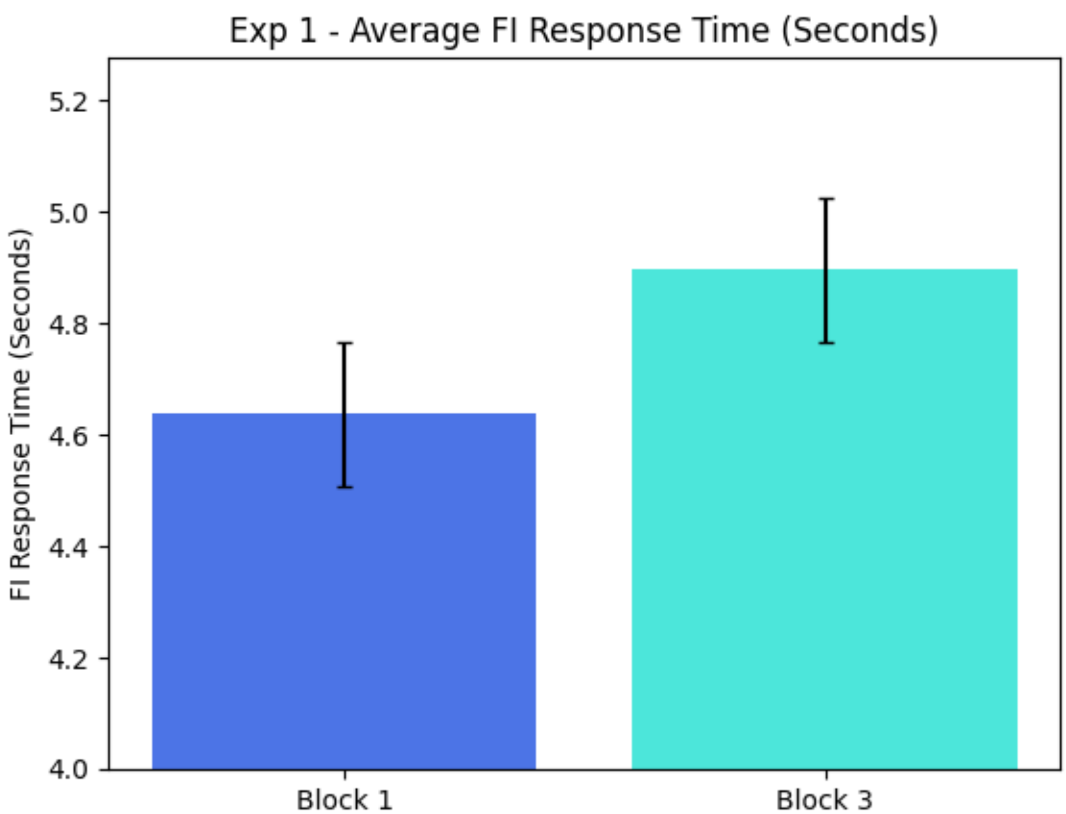

Figure 11: Average Response Time of FI Responses in Exp 1

5.5 Average Female-Intellectualized Choice Response Time in Exp. 2

The difference in the average female choice response time between Block 1 and Block 3 in Experiment 2 was not statistically significant for all participants ($t(30) = 0.49$, $p = 0.63$). As shown in Figure 12, the average response time for choosing female characters in Block 1 was 3.79 seconds, which slightly increased to 4.10 seconds in Block 3. This result suggests that there was no meaningful change in response time between Block 1 and Block 3 for female characters after the training phase in Block 2.

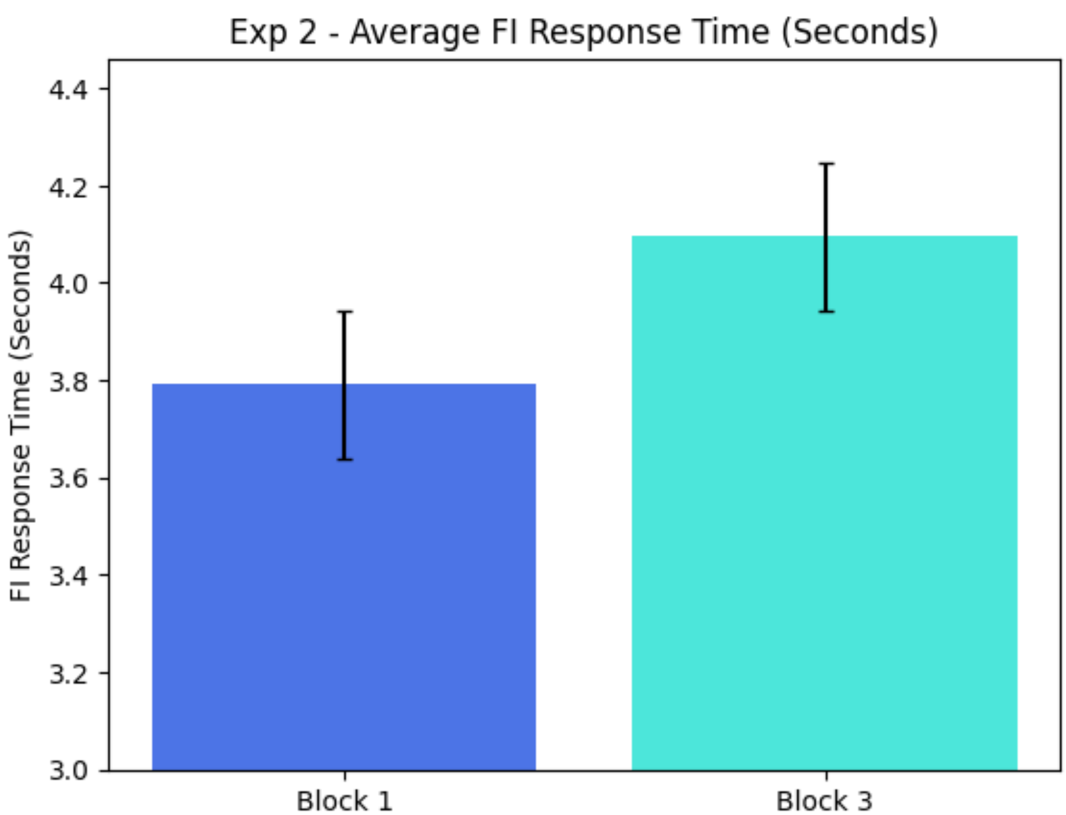

Figure 12: Average Response Time of FI Responses in Exp 2

5.6 Average MI Choice Response Time in Exp. 1

The difference in the average response choice numbers between Block 1 and Block 3 in Experiment 2 was not statistically significant for all participants ($t(30) = -0.34$, $p = 0.74$). As shown in Figure 13, the average response time in Block 1 was 5.27 seconds which decreased slightly to 5.08 seconds in Block 3 when participants were choosing male-intellectualized characters. This result indicates that there was no substantial change between Block 1 and Block 3 in the preference for male-intellectualized characters after the training phase in Block 2.

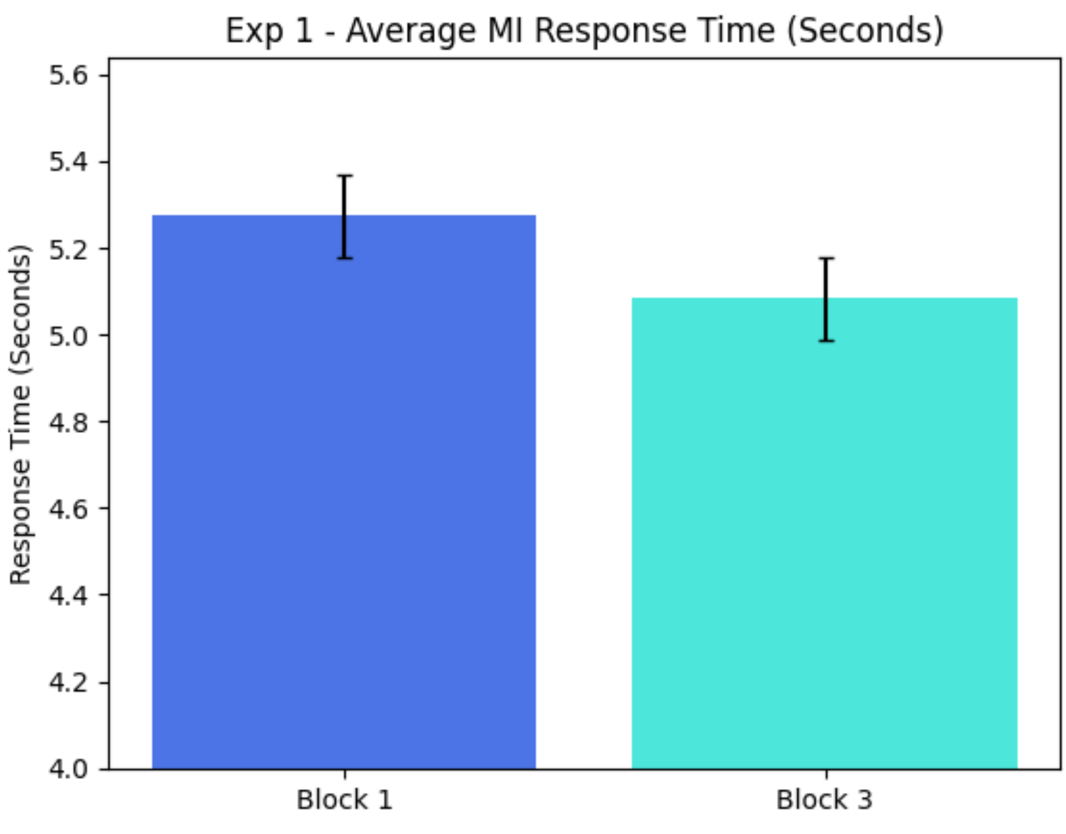

Figure 13: Average Response Time of FO/MI Responses in Exp 1

5.7 Average FO/MI Choice Response Time in Exp. 2

The difference in the average response time between Block 1 and Block 3 in Experiment 2 was not statistically significant for all participants ($t(30) = -0.44$, $p = 0.67$). As shown in Figure 14, the average response time for choosing male-intellectualized characters in Block 1 was 4.39 seconds, which slightly increased to 4.62 seconds in Block 3 when participants were choosing between female-only characters. This result indicates that there was no significant change in response time between Block 1 and Block 3 after the training phase in Block 2

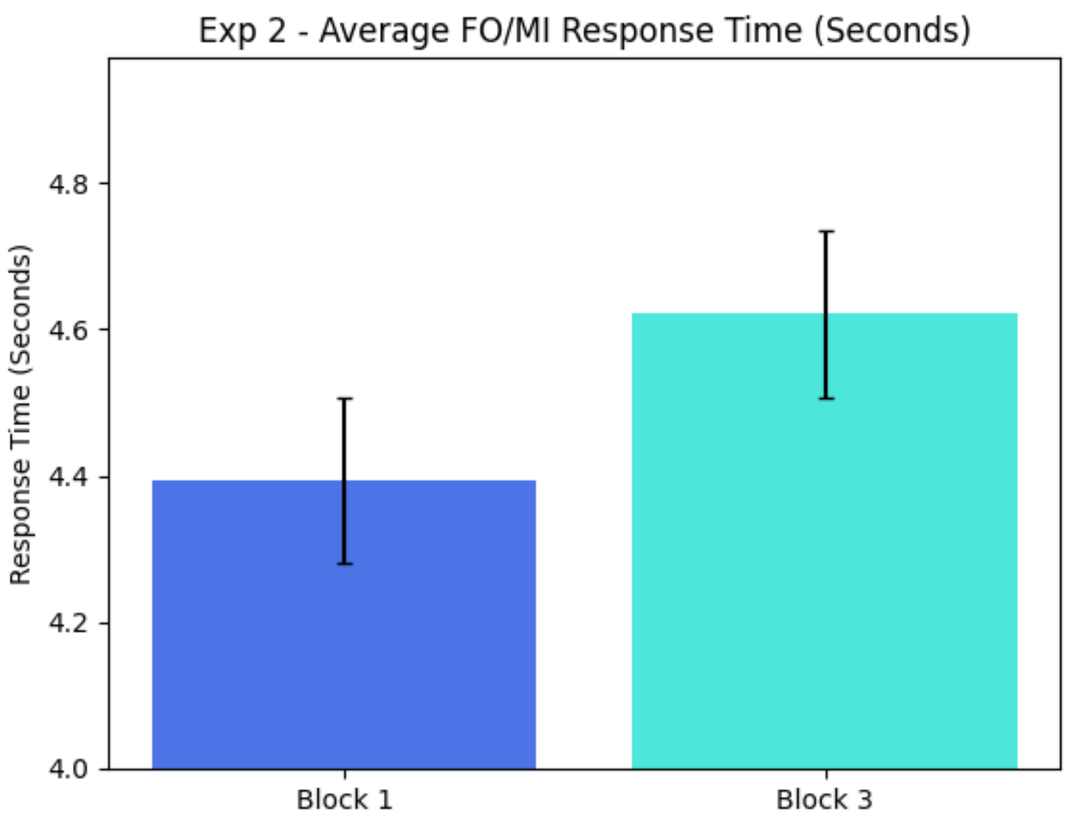

Figure 14: Average Response Time of FO/MI Responses in Exp 2

5.8 Average FI Pupil Size in Exp. 1

The difference in average pupil size between Block 1 and Block 3 in Experiment 1 was not statistically significant for all participants ($t(30) = -0.41$, $p = 0.68$). As shown in Figure 15, the average pupil size in Block 1 was 3.60 mm which increased slightly to 3.73 mm in Block 3 when participants were choosing between female-intellectualized characters. This result indicates that there was no significant change in pupil size between Block 1 and Block 3 after the training phase in Block 2.

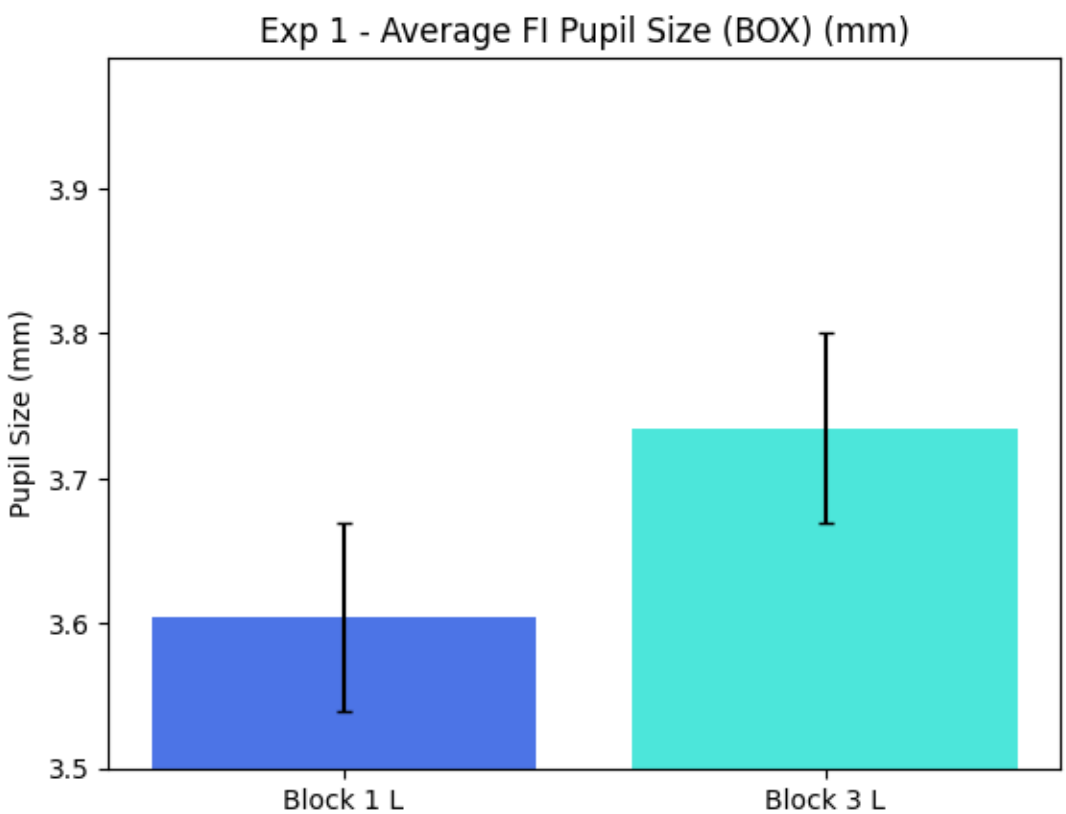

Figure 15: Average Pupil Size of FI Responses in Exp 1

5.9 Average FI Pupil Size in Exp. 2

The difference in average pupil size between Block 1 and Block 3 in Experiment 2 was not statistically significant for all participants ($t(30) = -1.03$, $p = 0.31$). As shown in Figure 16, the average pupil size in Block 1 was 3.26 mm, which increased to 3.65 mm in Block 3 when participants were choosing between female-intellectualized characters. This result indicates that there was no significant change in pupil size between Block 1 and Block 3 after the training phase in Block 2.

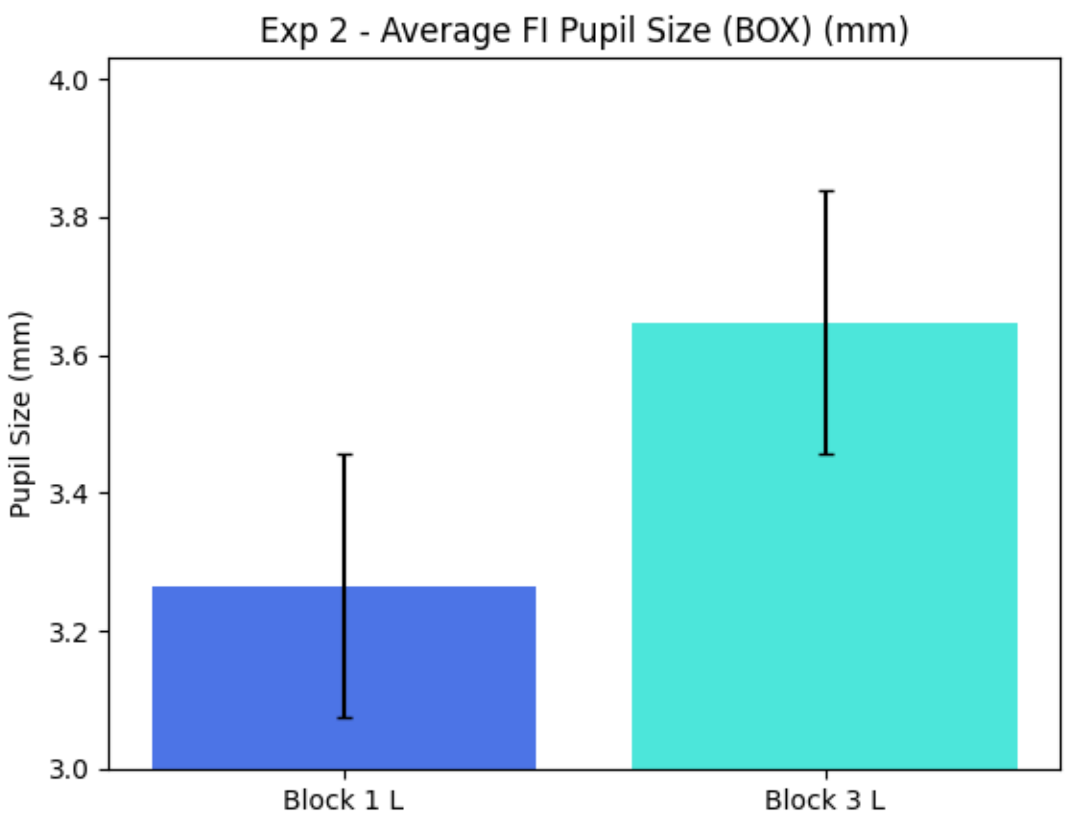

Figure 16: Average Pupil Size of FI Responses in Exp 2

5.10 Mean Pupil Size for All Trials

In Figure 17, the graph shows the pupil size for Experiment 1 is higher than that of Experiment 2. The difference in pupil size may suggest that participants in Experiment 1 were more engaged or experienced higher cognitive load while processing the information and making choices compared to participants in Experiment 2. Larger pupil sizes are often associated with increased mental effort, so it is possible that the stimuli in Experiment 1 required more cognitive processing from participants.

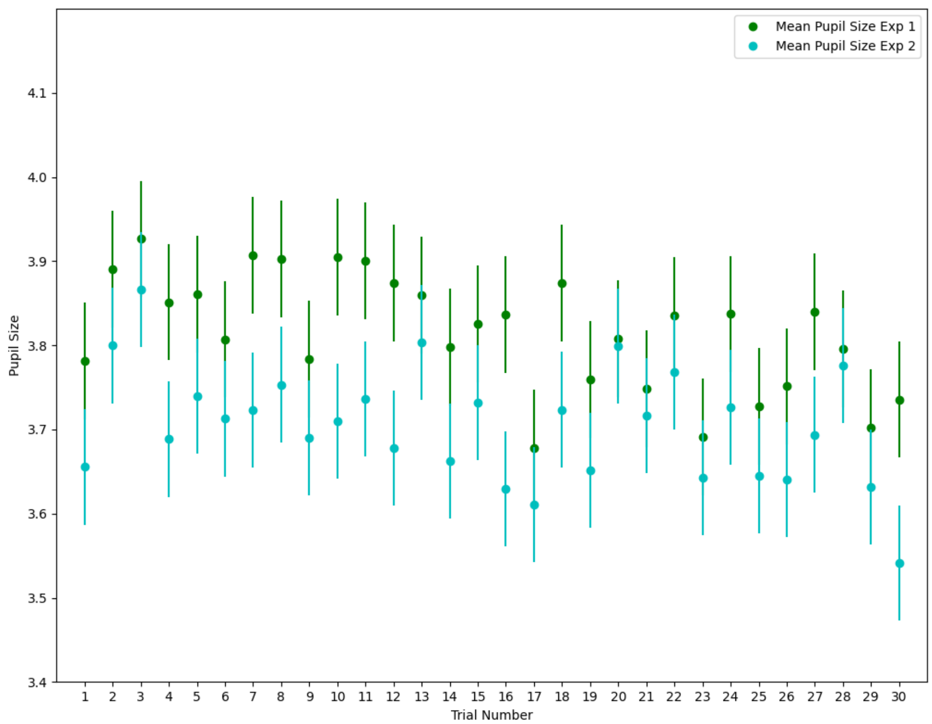

Figure 17: Mean Pupil Size for All Trials in Exp 1 and Exp 2

Although some of our results showed statistically significant differences between male and female participants, the majority of our analyses did not show a significant influence of gender on response times or response choice numbers. The overall pattern of our results does not provide strong evidence for a consistent gender effect in the context of our study. In Experiment 1, Block 1, 52.5% of all participants chose FI with a slight increase in Block 3 to 56.25% which showed no statistical significance ($t(30) = -0.68$, $p = 0.5$). Experiment 2, however, 30% of all participants chose FI with a significant increase to 76.87% in Block 3 which showed statistical significance($t(30) = -10.54$, $p < 0.005$).

The difference in average response time between male and female participants when choosing FI characters in Experiment 1, Block 1 was not statistically significant ($t(30) = -0.48$, $p = 0.64$). The average response time for male participants was 4.36 seconds while the average response times for female participants was 4.90 seconds. This result indicates that there was no significant influence on gender on the response times when choosing FI characters in Experiment 1, Block 1.

The difference in average response time between male and female participants when choosing FI characters in Experiment 1, Block 3 was not statistically significant ($t(30) = -0.47$, $p = 0.64$). The average response time for male participants was 4.64 seconds while the average response times for female participants was 5.16 seconds. This result indicates that there was no significant influence on gender on the response times when choosing FI characters in Experiment 1, Block 3.

The difference in average response time between male and female participants when choosing FI characters in Experiment 2, Block 1 was statistically significant ($t(30) = -2.51$, $p < 0.05$). The average response time for male participants was 2.613 seconds while the average response times for female participants was 4.971 seconds. This result indicates that there was some significant influence on gender on the response times when choosing FI characters in Experiment 2, Block 1.

The difference in average response time between male and female participants when choosing FI characters in Experiment 2, Block 3 was statistically significant ($t(30) = -2.21$, $p < 0.05$). The average response time for male participants was 3.51 seconds while the average response times for female participants was 4.67 seconds. This result indicates that there was some significant influence on gender on the response times when choosing FI characters in Experiment 2, Block 3.

The difference in average FI response choice numbers between male and female participants when choosing characters in Experiment 1, Block 1 was not statistically significant ($t(30) = -1.87$, $p = 0.08$). The average FI response choice numbers for male participants was 4.5 while the average FI response choice numbers for female participants was 6.0. This result indicates that there was no significant influence on gender on the FI response choices when choosing FI characters in Experiment 1, Block 1.

The difference in average FI response choice numbers between male and female participants when choosing characters in Experiment 1, Block 3 was not statistically significant ($t(30) = -0.36$, $p = 0.73$). The average FI response choice numbers for male participants was 5.5 while the average FI response choice numbers for female participants was 5.75. This result indicates that there was no significant influence on gender on the FI response choices when choosing FI characters in Experiment 1, Block 3.

The difference in average FI response choice numbers between male and female participants when choosing characters in Experiment 2, Block 1 was not statistically significant ($t(30) = -1.78$, $p = 0.1$). The average FI response choice numbers for male participants was 2.375 while the average FI response choice numbers for female participants was 3.625. This result indicates that there was no significant influence on gender on the FI response choices when choosing FI characters in Experiment 2, Block 1.

The difference in average FI response choice numbers between male and female participants when choosing characters in Experiment 2, Block 3 was not statistically significant ($t(30) = 0.26$, $p = 0.8$). The average FI response choice numbers for male participants was 7.75 while the average FI response choice numbers for female participants was 7.625. This result indicates that there was no significant influence on gender on the FI response choices when choosing FI characters in Experiment 2, Block 3.

The difference in average response time amongst all participants when choosing between FI and FO/MI choices in Experiment 1, Block 1 was not statistically significant ($t(30) = -0.72$, $p = 0.48$). The average response time for all participants that chose FI was 4.636 seconds while the average response time for all participants that chose FO/MI was 5.273 seconds. This result indicates that there was no significant influence on gender on the response times when choosing between FI and FO/MI characters in Experiment 1, Block 1.

The difference in average response time amongst all participants when choosing between FI and FO/MI choices in Experiment 1, Block 3 was not statistically significant ($t(30) = -0.25$, $p = 0.8$). The average response time for all participants that chose FI was 4.895 seconds while the average response time for all participants that chose FO/MI was 5.082 seconds. This result indicates that there was no significant influence on gender on the response times when choosing between FI and FO/MI characters in Experiment 1, Block 3.

The difference in average response time amongst all participants when choosing between FI and FO/MI choices in Experiment 2, Block 1 was not statistically significant ($t(30) = -0.9$, $p = 0.38$). The average response time for all participants that chose FI was 3.792 seconds while the average response time for all participants that chose FO/MI was 4.394 seconds. This result indicates that there was no significant influence on gender on the response times when choosing between FI and FO/MI characters in Experiment 2, Block 1.

The difference in average response time amongst all participants when choosing between FI and FO/MI choices in Experiment 2, Block 3 was not statistically significant ($t(30) = -1.16$, $p = 0.26$). The average response time for all participants that chose FI was 4.095 seconds while the average response time for all participants that chose FO/MI was 4.621 seconds. This result indicates that there was no significant influence on gender on the response times when choosing between FI and FO/MI characters in Experiment 2, Block 3.

5.11 General Discussion

In Experiment 1, the increase in female character choices between Blocks 1 and 3 confirms previous research findings, suggesting that exposure to same-gender characters in the training phase effectively reduces gender bias [3]. This finding supports the idea that gender bias can be mitigated through intervention.

The results of Experiment 2 also provide insight into gender bias and how effective exposure-based intervention is. More female-intellectualized responses in Experiment 2, Block 3 suggests that training on male characters contributed to the reduction of gender bias. This finding is significant because it indicates that exposure to both genders in a non-stereotypical setting might help give a more balanced picture of individuals.

Interestingly, in the graph in Fig. 16, the last trial has the lowest values of mean pupil size for both Experiment 1 and Experiment 2. This could indicate a decrease in cognitive load or engagement towards the end of the experiments. This could be due to a variety of factors, such as participants becoming more familiar with the tasks or experiencing fatigue as the trials went on.

The preference for female-intellectualized characters over female-only characters in Block 3 of Experiment 2 is significant because it demonstrates that the participants are less likely to associate women with stereotypical traits such as focusing solely on appearance. This result implies that exposure-based interventions can help challenge and mitigate gender bias.

VI. CONCLUSION

The results of the study support both hypotheses. In Experiment 1, our findings confirmed previous research [3], indicating that the participants chose a significantly greater number of female characters in Block 3 compared to Block 1 after the training intervention. This suggests that the training block was successful in promoting the selection of female characters, in turn reducing or untraining gender bias.

In Experiment 2, our results showed a significant increase in the proportion of female-intellectualized choices compared to female-only choices in Block 3 after the training block. This supports the alternative hypothesis (H1) and indicates that the exposure to male-intellectualized characters during the training block in Experiment 2 effectively influenced participants' choices, leading them to select more female-intellectualized characters over female-only characters. This finding demonstrates the potential of exposure-based interventions in altering perceptions and stereotypes related to gender roles.

In conclusion, the study highlights the effectiveness of exposure-based interventions in mitigating gender bias. By presenting participants with different characters, especially those challenging traditional stereotypes, individuals' perceptions can be shaped to reduce gender bias in their choices. These findings have important implications for the development of interventions aimed at promoting gender equality and reducing gender bias.

6.1 Future Work

Implementing a fixed head mount during the experiment would allow for more accurate and precise eye-tracking measurements by reducing head movement and ensuring a consistent viewing distance. Investigating different age groups could reveal the influence of age on gender bias and help identify whether biases change or persist across different age ranges. Additionally, conducting the experiment with a larger sample size would allow for more robust statistical analysis and increase the likelihood of identifying meaningful patterns in the data.

Integrating the decision tree proposed by Valtakari et al. as shown in Figure 18 into our eye-tracking experiment design could provide additional insights into creating a well thought out experiment. Ensuring a random order of character presentation could eliminate potential order effects and increase the validity of these findings. Expanding the number of trials could provide a more comprehensive assessment of gender bias. However, it is important to maintain participant engagement by incorporating breaks if trial numbers increase. Investigating the effects of male-intellectualized and male-only images could provide a more comprehensive understanding of gender bias in visual representations.

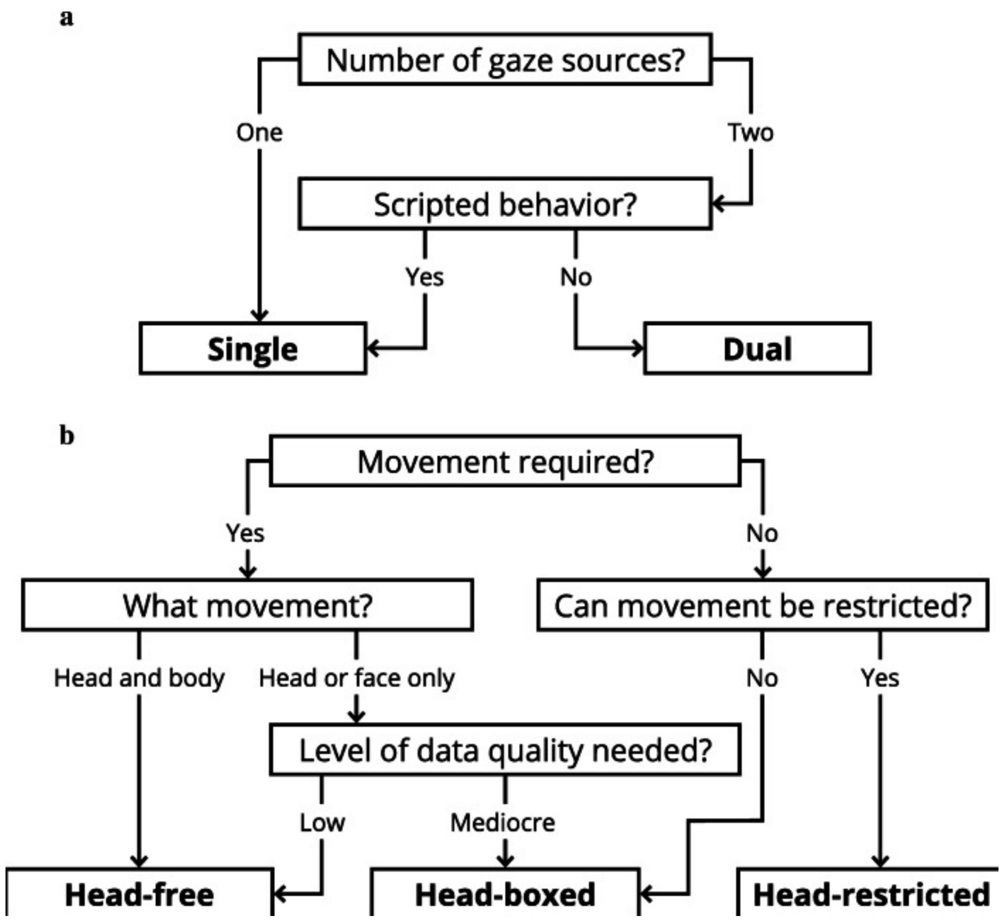

Figure 18: Valtakari et al. decision tree [29]

Introducing a control group that does not receive training sessions (Block 2 in both experiments) would help determine if the observed changes in female choices are attributable to training. Exploring the effects of variables such as character position on screen, size of characters, size of background images, and colors used in the images could be changed to provide further insights into the impact of visual elements on gender bias. By addressing these potential improvements and future research avenues, the study's design and outcomes can be further strengthened, contributing to a deeper understanding of gender bias and effective methods for mitigating it.

VII. REFERENCES

[1] England, P. *Theory on Gender / Feminism on Theory. (Social Institutions and Social Change)*. Aldine De Gruyter, 1993.

[2] Svetlana Bogomolova, Harmen Oppewal, Justin Cohen, and Jun Yao. "How the Layout of a Unit Price Label Affects Eye-Movements and Product Choice: An Eye-Tracking Investigation." Journal of Business Research, vol. 111, 2020, pp. 102–16, https://doi.org/10.1016/j.jbusres.2018.10.049.

[3] Reem Albaghli, Chaz Chang, Sarah Almahmeed, and Nada Attar. 2022. Untraining Ethnocentric Biases about Gender Roles: A Preliminary Empirical Study Presenting Art as Stimulus. In Proceedings of Mensch und Computer 2022 (MuC '22). Association for Computing Machinery, New York, NY, USA, 376–381. https://doi.org/10.1145/3543758.3547543

[4] Jenna Cryan, Shiliang Tang, Xinyi Zhang, Miriam Metzger, Haitao Zheng, and Ben Y. Zhao. "Detecting Gender Stereotypes: Lexicon Vs. Supervised Learning Methods." Conference on Human Factors in Computing Systems - Proceedings, ACM, 2020, pp. 1–11, https://doi.org/10.1145/3313831.3376488.

[5] Andrich, Aliya, and Emese Domahidi. "A Leader and a Lady? A Computational Approach to Detection of Political Gender Stereotypes in Facebook User Comments." International Journal of Communication (Online), vol. 17, 2023, p. 236–.

[6] Sean Matthews, John Hudzina, Dawn Sepehr. Gender and Racial Stereotype Detection in Legal Opinion Word Embeddings. 2022, https://doi.org/10.48550/arxiv.2203.13369.

[7] Brooke, S. J. "Trouble in Programmer's Paradise: Gender-Biases in Sharing and Recognising Technical Knowledge on Stack Overflow." Information, Communication & Society, vol. 24, no. 14, 2021, pp. 2091–112, https://doi.org/10.1080/1369118X.2021.1962943.

[8] Zyte. (2023, May 10). Gender Inequality Across Programming Languages. Retrieved from https://www.zyte.com/blog/gender-inequality-across-programming-languages/

[9] BBC. (2016, February 15). Github coding study suggests gender bias. Retrieved from https://www.bbc.com/news/technology-35559439

[10] Liu, Haochen. Detecting and Mitigating Bias in Natural Languages. ProQuest Dissertations Publishing, 2022.

[11] Corinne A. Moss-Racusin, Christina Sanzari, Nava Caluori, and Helena Rabasco. "Gender Bias Produces Gender Gaps in STEM Engagement." Sex Roles, vol. 79, no. 11-12, 2018, pp. 651–70, https://doi.org/10.1007/s11199-018-0902-z.

[12] Kirtane, Neeraja, and Tanvi Anand. Mitigating Gender Stereotypes in Hindi and Marathi. 2022, https://doi.org/10.48550/arxiv.2205.05901.

[13] Kevin Beltran, Cody Rowland, Nicki Hashemi, Anh Nguyen, Lane Harrison, Sophie Engle, and Beste F Yuksel. 2021. Reducing Implicit Gender Bias Using a Virtual Workplace Environment. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, Article 277, 1–7. https://doi.org/10.1145/3411763.3451739.

[14] Niilo V. Valtakari, Ignace T. C. Hooge, Charlotte Viktorsson, Pär Nyström, Terje Falck-Ytter, and Roy S. Hessels. "Eye Tracking in Human Interaction: Possibilities and Limitations." Behavior Research Methods, vol. 53, no. 4, 2021, pp. 1592–608, https://doi.org/10.3758/s13428-020-01517-x.

[15] Jess Kerr-Gaffney, Amy Harrison, and Kate Tchanturia. "Eye-tracking Research in Eating Disorders: A Systematic Review." The International Journal of Eating Disorders, vol. 52, no. 1, 2019, pp. 3–27, https://doi.org/10.1002/eat.22998.

[16] Jean Pfiffelmann, Nathalie Dens, and Sébastien Soulez. “Personalized Advertisements with Integration of Names and Photographs: An Eye-Tracking Experiment.” Journal of Business Research, vol. 111, no. 111, 2020, pp. 196–207, https://doi.org/10.1016/j.jbusres.2019.08.017.

[17] Płużyczka, Monika. “The First Hundred Years: a History of Eye Tracking as a Research Method.” Applied Linguistics Papers, vol. 4/2018, no. 25, 2018, pp. 101–16, https://doi.org/10.32612/uw.25449354.2018.4.pp.101-116.

[18] Holmqvist, Kenneth, and Richard Andersson. *Eye Tracking: A Comprehensive Guide to Methods, Paradigms and Measures*. Lund Eye-Tracking Research Institute, 2018.

[19] Krzysztof Krejtz, Andrew T. Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. “Eye Tracking Cognitive Load Using Pupil Diameter and Microsaccades with Fixed Gaze.” PloS One, vol. 13, no. 9, 2018, pp. e0203629–e0203629, https://doi.org/10.1371/journal.pone.0203629.

[20] Dossa, F., and N. N. Baxter. “Reducing Gender Bias in Surgery.” British Journal of Surgery, vol. 105, no. 13, 2018, pp. 1707–09, https://doi.org/10.1002/bjs.11042.

[21] Natalia Sevcenko, Tobias Appel, Manuel Ninaus, Korbinian Moeller, and Peter Gerjets. “Theory-Based Approach for Assessing Cognitive Load During Time-Critical Resource-Managing Human–computer Interactions: An Eye-Tracking Study.” Journal on Multimodal User Interfaces, vol. 17, no. 1, 2023, pp. 1–19, https://doi.org/10.1007/s12193-022-00398-y.

[22] Man, Tsoey Wun, and Peter J. Hills. “Eye-Tracking the Own-Gender Bias in Face Recognition: Other-Gender Faces Are Viewed Differently to Own-Gender Faces.” Visual Cognition, vol. 24, no. 9-10, 2016, pp. 447–58, https://doi.org/10.1080/13506285.2017.1301614.

[23] Kawakami, K., Williams, A., Sidhu, D., Choma, B. L., Rodriguez-Bailón, R., Cañadas, E., Chung, D., and Hugenberg, K. “An Eye for the I: Preferential Attention to the Eyes of

Ingroup Members.” Journal of Personality and Social Psychology, vol. 107, no. 1, 2014, pp. 1–20, https://doi.org/10.1037/a0036838.

[24] Wade, Nicholas J., and Benjamin W. Tatler. “Did Javal Measure Eye Movements During Reading?” Journal of Eye Movement Research, vol. 2, no. 5, 2009,

[25] Wilbanks, Bryan A., Aroke, Edwin; Dudding, Katherine M. “Using Eye Tracking for Measuring Cognitive Workload During Clinical Simulations: Literature Review and Synthesis.” Computers, Informatics, Nursing, vol. 39, no. 9, 2021, pp. 499–507, https://doi.org/10.1097/CIN.0000000000000704.

https://doi.org/10.1016/j.jesp.2022.104406.

[26] SR Labs. (n.d.) Tobii Pro Fusion. Retrieeved from https://www.srlabs.it/en/scientific-research/hardware-products/tobii-pro-fusion/

[27] Tobias B. Konitzer, Shanto Iyengar, Nicholas A. Valentino, Stuart Soroka, Raymond M. Duch. (2019) Ethnocentrism versus group-specific stereotyping in immigration opinion: cross-national evidence on the distinctiveness of immigrant groups. *Journal of Ethnic and Migration Studies* 45:7, pages 1051-1074.

[28] Wirth, Benedikt Emanuel, and Dirk Wentura. “Not Lie Detection but Stereotypes: Response Priming Reveals a Gender Bias in Facial Trustworthiness Evaluations.” Journal of Experimental Social Psychology, vol. 104, 2023, p. 104406–,

[29] Leuschner, Anna. “Why So Low?: On Indirect Effects of Gender Bias in Philosophy.” Metaphilosophy, vol. 50, no. 3, 2019, pp. 231–49, https://doi.org/10.1111/meta.12353.

https://doi.org/10.16910/jemr.2.5.5.

[30] Mayfair Gallery. (n.d.) Orientalist Paintings: 19th Century Fantasies of the East. Retrieved from https://www.mayfairgallery.com/blog/orientalist-paintings-19th-century-fantasies-east

[31] Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. “To ‘See’ Is to Stereotype: Image Tagging Algorithms, Gender Recognition, and the Accuracy-Fairness Trade-Off.” Proceedings of the ACM on Human-Computer Interaction, vol. 4, no. CSCW3, 2021, pp. 1–31, https://doi.org/10.1145/3432931.

[32] Debue, Nicolas, and Ce´cile van de Leemput. “What Does Germane Load Mean? An Empirical Contribution to the Cognitive Load Theory.” Frontiers in Psychology, vol. 5, 2014, pp. 1099–1099, https://doi.org/10.3389/fpsyg.2014.01099.

[33] Eiman A. Eissa, Hala A. Guta, and Rana S. Hassan. “Representations of Arab Women in Hollywood Pre- and Post- 9/11.” Journal of International Women’s Studies, vol. 24, no. 5, 2022, pp. 1–20.

[34] Peterson, Biederman, L. A., Andersen, D., Ditonto, T. M., & Roe, K. (2019). Mitigating gender bias in student evaluations of teaching. PloS One, 14(5), e0216241–e0216241. https://doi.org/10.1371/journal.pone.0216241

[35] Siegenthaler E, Costela FM, McCamy MB, Di Stasi LL, Otero-Millan J, Sonderegger A, et al. Task difficulty in mental arithmetic affects microsaccadic rates and magnitudes. European Journal of Neuroscience. 2014; 39(1):1–8.

[36] Skulmowski, Alexander, and Kate Man Xu. “Understanding Cognitive Load in Digital and Online Learning: a New Perspective on Extraneous Cognitive Load.” Educational Psychology Review, vol. 34, no. 1, 2022, pp. 171–96, https://doi.org/10.1007/s10648-021-09624-7.

[37] Martin, Wendy. “Beyond Intellectual Sexism: A New Woman, a New Reality . Joan I. Roberts New Research on Women and Sex Roles at the University of Michigan . Dorothy G. McGuigan.” Signs: Journal of Women in Culture and Society 3, no. 2 (1977): 497–499.

[38] Christie's (n.d.). Orientalist Art: A Collecting Guide. Retrieved from

https://www.christies.com/features/Orientalist-Art-Collecting-guide-8426-1.aspx