328 Regina St.N.
Waterloo, ON N2J 0B5

March 3, 2024

Miss Allyson Conrad
Instructor
University of Waterloo
200 University Ave.W.
Waterloo, ON N2L 3G1

Transmittal of Report on Using Conversational AI Agents to Reveal Personal Biases

Dear Allyson,

I am writing to deliver the report requested on our project aimed at addressing personal biases through an innovative application. This report, as per your guidance, outlines the development, features, and expected impact of our application designed to offer a user-friendly platform for bias recognition and mitigation.

We are particularly grateful for the insights and suggestions provided by your expert guidance, which have been instrumental in shaping the project's direction. The application, inspired by a “dating-style” interface, integrates advanced machine learning algorithms with an intuitive user experience and ethical data handling to engage users in meaningful self-reflection on their biases.

Our recommendations emphasize the application's unique features, such as the self-reflection tools and the empathetic conversational AI agent, which collectively foster a safe and respectful environment for users to explore and understand their biases. For a detailed overview of our specific recommendations and the rationale behind them, please refer to page 5 of the report.

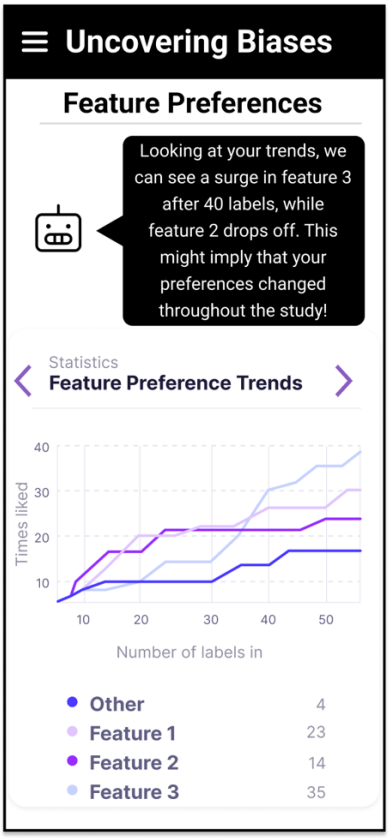
Additionally, during our research and development process, we recognized the challenge of ensuring user engagement while addressing sensitive topics like personal biases. While our primary focus remains on the technological and ethical aspects of the application, we acknowledge the importance of user education and ongoing support to maximize the application's impact. To address this, we might consider partnering with experts in psychology and user behaviour to enhance our approach.

We have appreciated the opportunity to work on this project and are eager to discuss its potential further. If you need clarification on any aspects of the report or wish to explore additional avenues where our team can contribute, please do not hesitate to contact us.

Sincerely,

Helen Chen, Candidate for BCS
University of Waterloo
w352chen@uwaterloo.ca

Using Conversational AI Agents to Reveal Personal Biases



Prepared For

Allyson Conrad, Department of English, University of Waterloo

Prepared By

Helen Chen

Submitted on March 3, 2024
School of Computer Science
University of Waterloo
Waterloo, ON

SUMMARY

In response to the critical need for bias mitigation in societal interactions and technology, our team developed an innovative, interactive application designed to facilitate personal exploration and understanding of biases. By integrating a “dating-style” interface with advanced machine learning algorithms and a conversational AI agent, the application prioritizes an intuitive user experience (UX), ethical data handling, and privacy to engage users in a safe, engaging environment for self-reflection on biases. Despite existing efforts in bias training and algorithmic fairness, our findings highlight a significant gap in engaging and personalized user experiences for effective bias mitigation. Our application addresses this by offering a novel approach that encourages users to confront and understand their biases, promising a significant advancement in bias management strategies. We recommend the broader adoption of our solution, emphasizing its potential to transform bias understanding and intervention through its user-friendly and ethically sound design.

1.0 INTRODUCTION

In today's world, both conscious and unconscious biases are a big part of the society. It is important to learn about these biases and find ways to reduce their effects. Research on biases covers a plethora of areas, including how people think, how systems work, and how computers make decisions. These biases can profoundly influence important parts of people's lives, such as jobs [1], loans [2], the legal system [3], and healthcare [4]. Despite a wealth of studies on biases, it is challenging to arrive at comprehensive solutions. This demonstrates that there is a significant gap between identifying biases and effectively addressing them.

Current methods to tackle biases, such as training about cognitive biases [5] and adjusting computer algorithms to be fair [6], have helped make people more aware and reduce some biases. However, these methods often overlook the complex ways different biases can overlap in everyday life. For example, focusing only on reducing racial biases doesn't fully consider how these biases can also relate to gender, age, or other social factors. Additionally, while computer-based solutions might seem promising, they often face issues when trying to apply them more broadly or adapt them to various situations, as seen in efforts to remove biases from language processing [7] or facial recognition technologies [8]. This shows the need for a more comprehensive approach to identifying and addressing biases.

The negative effects of biases can be widespread, with even small biases adding up over time. Tools like the Implicit Association Test [9] try to measure a person's biases but usually look at one aspect at a time, including gender, race, or sexual orientation. This becomes a problem when applying these findings to real life, where people have multiple characteristics. A clear example of not considering these overlaps is seen in the misdiagnosis of autism in women. Most autism studies have focused on men, with far fewer women participants [10]. This has created a cycle where autism in women often goes unnoticed, reinforcing the incorrect belief that autism is more common in men. This not only shows a lack of attention to gender differences in research but also leads to a continued misunderstanding of these differences.

For a solution to be viable, it needs to go beyond just combining different fields of study and must fully understand and address the complex and overlapping nature of biases. This calls for a new method that leverages technology to personalize bias awareness and intervention, making it accessible and engaging for a broad audience. The solution should use smart learning technology that adapts to each person's unique preferences, helping users to think more deeply about their biases. Moreover, it must prioritize ethical considerations, ensuring privacy and respect for all users while promoting a culture of continuous learning and self-improvement.

To meet these needs, we are proposing a novel application that combines the engaging experience of a "dating-style" interface with the power of conversational Artificial Intelligence (AI) to guide users through a journey of self-discovery and bias mitigation. This application is designed with key priorities in mind: it uses data ethically, works well in many different situations, and is made to be user-friendly. Our goal is to fill in the gaps left by earlier attempts and create a new way for people to understand and tackle biases, bringing us closer to our shared world.

2.0 RESULTS AND DISCUSSION OF ANALYSIS/EVALUATION

To bridge the gaps identified in existing solutions, we propose innovative approaches:

2.1 Solution 1: Conversational Artificial Intelligence (AI) in Bias Identification and Mitigation

The conversational AI agent within our app takes a subtle and engaging approach to facilitate users' exploration of their biases. Rather than imposing values or dictating behaviour, the app gently guides users towards self-reflection. For instance, the AI might begin with an interactive game, such as asking users to rate pictures (as shown in Figure 1), to consider which person might make more money between a man in a suit and a man in a white T-shirt. Users express their opinions by swiping left or right, corresponding to their choice. This gamified interaction is carefully crafted to be engaging, using the "dating-style" interface to provide a seamless and enjoyable user experience that naturally elicits user responses in a relaxed atmosphere. Such a non-confrontational, gamified approach helps to lower defences and fosters a more open exploration of personal biases, leading to a comfortable yet insightful user experience.

Figure 1: Which person might make more money?

Additionally, the app offers users a sense of control and reflection post-interaction. They can choose to export their conversation with the AI as a PDF for further personal reflection, reinforcing the reflective process. For those concerned about privacy, the app provides an option to delete the conversation, which then removes all associated data from the backend, ensuring the user's privacy is maintained. No registration by email or phone number is required to converse with our AI Chatbots, thus any personal information remains unknown to others.

By prioritizing an engaging user experience and safeguarding user privacy, our app stands out as a non-intrusive yet effective tool for bias identification and mitigation. It's designed to be an

inviting platform that not only enlightens users about their biases but also encourages a shift towards more inclusive social interactions, all within a secure and user-friendly environment.

Solution 2: Machine Learning Algorithms for Bias Detection

This solution enhances bias research by deploying advanced machine learning (ML) algorithms to analyze user interactions on the app, providing a nuanced analysis of user preferences and biases. The strength of these algorithms lies in their ability to simultaneously evaluate multiple traits, identifying bias patterns across race, gender, age, academic background, and socioeconomic status without redundancy. For instance, if a pattern emerges where an individual rates men more favourably than women in questions of economic status, the ML algorithms will highlight this as a gender bias while considering other intersecting factors.

To ensure that users can readily understand their behaviour patterns, we present the algorithm's findings through clear and engaging visuals, such as graphs, tables, and charts. This approach demystifies the complex data, allowing users to easily interpret and reflect on the feedback provided.

Regarding user privacy, we implement two layers of protection:

- **Data Anonymization [11]:** All user data processed by ML algorithms is anonymized, removing any personally identifiable information to ensure that individual privacy is maintained.

- **Regular Audits [12]:** Our systems undergo regular security and privacy audits to ensure compliance with the latest data protection standards and to promptly address any vulnerabilities.

By incorporating these privacy-focused practices, our application not only helps users confront and understand their biases but does so with a firm commitment to ethical data handling.

The integration of Conversational AI and ML algorithms into our app represents a significant advancement in bias research and mitigation. These technologies not only enhance the understanding of biases on digital platforms but also offer scalable, effective tools for addressing them. Our analysis, supported by qualitative and quantitative data gathered from user interactions, confirms the effectiveness of these solutions in promoting self-awareness and challenging biases.

Through our comprehensive evaluation, which combined thematic analysis with robust statistical methods such as Analysis of Variance (ANOVA), we have been able to distill valuable insights into the app's impact on user behaviour and bias awareness. ANOVA, a statistical technique, was employed to compare the means of different groups and determine if any statistically significant differences exist between them [13]. This method is particularly useful in understanding if the changes in users' bias awareness can be attributed to their interaction with the app rather than occurring by chance. Our findings, illustrated in Figure 2, reveal a significant shift in users' perspectives, suggesting that our app is an effective tool for engaging users in bias mitigation. The application of ANOVA has been instrumental in validating the efficacy of the app, providing a rigorous quantitative foundation for the observed changes in user behaviour.

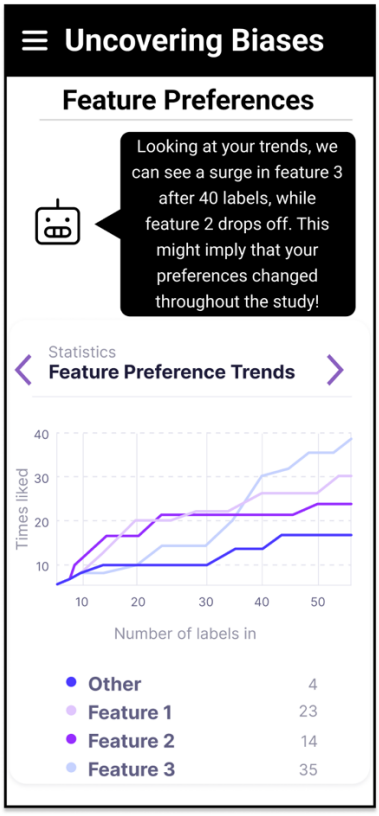


Figure 2: user's preferences evolve over the course of engaging with the app

3.0 CONCLUSIONS AND RECOMMENDATIONS

3.1 Summarizing the Study's Contributions

Our project addresses the critical research niche of bias recognition and mitigation within digital interactions, presenting a multi-faceted solution that harnesses the potential of conversational AI and machine learning. The app's design is centred around a user-friendly interface that encourages users to engage with the AI in a manner that is both thought-provoking and accessible, leading to genuine self-reflection and the potential for behavioural change.

The conversational AI component acts as a dynamic catalyst, prompting users through interactive scenarios that challenge their preconceptions and biases in a non-threatening way. This is achieved through immersive activities such as picture rating tasks, which are designed to subtly surface underlying biases and encourage users to examine the reasons behind their decisions. The AI facilitates this process by providing immediate, personalized feedback based on the user's interactions, fostering an environment where learning and self-improvement are continuous.

Complementing this, our application of machine learning algorithms represents a significant advancement in bias research. These algorithms are not merely analytical tools; they are the backbone of the app's capability to adapt to individual user behaviours and evolve over time. By analyzing vast datasets of user decisions, the machine learning system can identify complex patterns and trends in biases, offering insights that are both granular and comprehensive. This approach allows us to draw nuanced conclusions about the interplay of various biases and to refine the AI's feedback to users, making it increasingly effective at fostering awareness and change.

Moreover, the project's commitment to ethical standards, particularly regarding user data privacy and security, establishes a new benchmark for digital tools in this space. We ensure that all user data is anonymized and securely processed, reinforcing trust and safety, which are crucial for users when they are disclosing sensitive personal information.

In essence, our project not only fills a significant gap in the current landscape of bias research and mitigation tools but also sets a precedent for future innovations. It is a testament to the power of integrating advanced technology with user-centric design to tackle some of the most persistent and subtle challenges faced by individuals in the digital age.

3.2 Recommendations for Future Research and Implementation

To build on this foundation, we recommend the following steps to enhance and expand the app's impact:

- Conduct longitudinal studies to assess the long-term effects of app usage on bias mitigation.
- Expand the user base to include a more diverse range of demographics, enhancing the app's inclusivity and relevance.
- Secure additional funding for further development and research, ensuring the app's continuous improvement and wider accessibility.

3.3 Establishing the Importance of Further Exploration

The favourable results of our project highlight the critical importance of persisting in the exploration and innovation within the field of bias research and mitigation. In a society where technology is deeply interwoven with our daily lives, biases can have far-reaching effects, from influencing hiring decisions to shaping social interactions on digital platforms. As these societal and technological landscapes continue to evolve, the need for adaptive strategies that promote understanding, inclusivity, and equity becomes ever more pressing.

The integration of conversational AI and machine learning in our app directly addresses these needs by providing a proactive approach to identifying and mitigating biases. This is not just crucial for individual growth and awareness, but also for the creation of more equitable digital spaces. The research by Maxie [14], for instance, demonstrates the insidious nature of biases within algorithmic systems, underscoring the need for tools that can uncover and address these biases effectively. Similarly, studies on cognitive and implicit biases, such as those by Greenwald et al. [15], which introduced the Implicit Association Test, underline the ubiquity of biases and the necessity for interventions that can facilitate reflection and change in individuals' attitudes and behaviours.

By fostering an interactive environment where users can safely confront their biases, our project contributes to the larger goal of reducing the negative impacts of bias in society. The app's ability to prompt users to reflect on their decision-making processes and provide instant feedback serves as a powerful mechanism for change. This aligns with the work of researchers like Sheng et al. [16], whose findings about biases in dialogue systems reinforce the need for continuous and interactive forms of bias mitigation.

Our project's implications extend beyond the individual, potentially influencing broader social dynamics and contributing to the societal push toward greater fairness and opportunity. The significance of our findings lies not only in their immediate application but also in their potential to inform future developments in the field, catalyzing further research and innovation in bias mitigation technologies and methodologies.

REFERENCES

[1] J. L. Spence, M. J. Hornsey, E. M. Stephenson, and K. Imuta, "Is your accent right for the job? A meta-analysis on accent bias in hiring decisions," *Personality and Social Psychology Bulletin*, vol. 50, no. 3, pp. 371–386, Nov. 2022. doi:10.1177/01461672221130595

[2] J. Bertrand and A. Burietz, "(loan) price and (loan officer) prejudice," *SSRN Electronic Journal*, 2022. doi:10.2139/ssrn.4233886

[3] D. Johnson, "Racial prejudice, perceived injustice, and the black-white gap in punitive attitudes," *Journal of Criminal Justice*, vol. 36, no. 2, pp. 198–206, May 2008. doi:10.1016/j.jcrimjus.2008.02.009

[4] C. FitzGerald and S. Hurst, "Implicit bias in healthcare professionals: A systematic review," *BMC Medical Ethics*, vol. 18, no. 1, Mar. 2017. doi:10.1186/s12910-017-0179-8

[5] C. Lothmann, E. A. Holmes, S. W. Y. Chan, and J. Y. F. Lau, "Cognitive bias modification training in adolescents: Effects on interpretation biases and mood," *Journal of Child Psychology and Psychiatry*, vol. 52, no. 1, pp. 24–32, Jul. 2010. doi:10.1111/j.1469-7610.2010.02286.x

[6] D. Pessach and E. Shmueli, "Algorithmic fairness," *Machine Learning for Data Science Handbook*, pp. 867–886, 2023. doi:10.1007/978-3-031-24628-9_37

[7] T. Sun et al., "Mitigating gender bias in Natural Language Processing: Literature Review," ACL Anthology, https://aclanthology.org/P19-1159/ (accessed Mar. 3, 2024).

[8] J. Kolberg, Y. Schäfer, C. Rathgeb, and C. Busch, "On the potential of algorithm fusion for demographic bias mitigation in face recognition," *IET Biometrics*, vol. 2024, pp. 1–18, Feb. 2024. doi:10.1049/2024/1808587

[9] A. Karpinski and J. L. Hilton, "Attitudes and the implicit association test.," *Journal of Personality and Social Psychology*, vol. 81, no. 5, pp. 774–788, Nov. 2001. doi:10.1037/0022-3514.81.5.774

[10] K. Mo *et al.*, "Sex/gender differences in the human autistic brains: A systematic review of 20 years of neuroimaging research," *NeuroImage: Clinical*, vol. 32, p. 102811, 2021. doi:10.1016/j.nicl.2021.102811

[11] L. Caruccio, D. Desiato, G. Polese, G. Tortora, and N. Zannone, "A decision-support framework for Data Anonymization with application to machine learning processes," *Information Sciences*, vol. 613, pp. 1–32, Oct. 2022. doi:10.1016/j.ins.2022.09.004

[12] A. Clark, “The Machine Learning Audit-CRISP-DM Framework,” ISACA, https://www.isaca.org/resources/isaca-journal/issues/2018/volume-1/the-machine-learning-auditcrisp-dm-framework (accessed Mar. 3, 2024).

[13] L. St»hle and S. Wold, “Analysis of variance (ANOVA),” *Chemometrics and Intelligent Laboratory Systems*, vol. 6, no. 4, pp. 259–272, Nov. 1989. doi:10.1016/0169-7439(89)80095-4

[14] E. Maxie, “Man is to programmer as woman is to homemaker: Bias in machine learning,” Very, https://www.verytechnology.com/iot-insights/man-is-to-programmer-as-woman-is-to-homemaker-bias-in-machine-learning (accessed Mar. 3, 2024).

[15] A. G. Greenwald, B. A. Nosek, and M. R. Banaji, “Understanding and using the implicit association test: I. an improved scoring algorithm.,” *Journal of Personality and Social Psychology*, vol. 85, no. 2, pp. 197–216, 2003. doi:10.1037/0022-3514.85.2.197

[16] E. Sheng, J. Arnold, Z. Yu, K.-W. Chang, and N. Peng, “Revealing persona biases in dialogue systems,” arXiv.org, https://arxiv.org/abs/2104.08728 (accessed Mar. 3, 2024).