

# **Refu-AI Selection and Ethical Assessment: A Critical Review of ChatGPT's Capabilities(1484 words)**

## **Introduction**

The Australian government is considering the implementation of an AI system, Refu-AI, aimed at streamlining the processing of refugee protection visa applications. The system, trained on the data from refugee cases and the conditions in their countries of origin, offers a confidence rating on applicants' qualifications as refugees, thereby assisting human case workers in their decision-making. However, the choice between two versions - one prioritizing accuracy with limited explanation and another offering comprehensive output explanations with a potential trade-off on accuracy - underscores the complexity inherent in integrating AI solutions into ethically charged, high-stakes environments such as asylum grant processes.

This research essay is composed of three parts. First, it critiques the overstated ability of ChatGPT in selecting between versions of Refu-AI. Secondly, considering the imperative of fairness in this context and the potential risks posed by the explainability of biased models, it recommends Version One to the government. Lastly, combining the performance of ChatGPT in case analysis with the distinctive characteristics of moral reasoning, it asserts that ChatGPT currently retains limitations in conducting ethical assessments of AI tools.

## **1. Discussion on the recommendation of ChatGPT**

To examine this case in detail, I have input all the relevant information about the case and respective features of the two models into ChatGPT. Based on the context, ChatGPT has determined that the government should purchase Version One. Here are the four reasons for this decision: 1. Version One

minimizes the risk of gender disparities in the application process.2. Version One has higher accuracy for specific countries.3. Version One ensures privacy and security.4. Version One upholds human decision-making. This section of the essay analyzes these points in turn.

ChatGPT's first two arguments are sound and logically coherent. Compared to Version Two, Version One's superior performance in terms of accuracy can minimize the risk of unjustly rejecting legitimate refugee claims. Furthermore, Version One's lesser discrepancy in predictive outcomes for applicants of different genders could help mitigate the risk of discrimination based on the country of origin. Conversely, bias in AI models constitutes a pressing issue that elicits significant concerns in today's world. If a government were to deploy biased models, it could precipitate ineffective decision-making and reduced productivity, while concurrently tarnishing the government's reputation [1]. Therefore, Version One is more suitable for evaluating eligibility for protection visas in this context.

The third major point made by ChatGPT is that, due to the lack of explanatory reports, Version One could better protect applicant confidentiality, making it a more suitable choice for refugee visa adjudication. However, this assessment is insufficiently comprehensive. While it is true that Version One does not offer explanatory reports, this does not inherently signify it ensures confidentiality. Moreover, with the deployment of privacy-preserving machine learning (PPML) and anonymization techniques [2], we can assure data privacy and security while maintaining the model's explainability. ChatGPT has not furnished compelling evidence to argue that the heightened interpretability of Version Two would lead to privacy breaches. Furthermore, it failed to acknowledge a critical piece of information from the case: before sharing related data with other countries, the outputs from Version Two can undergo anonymization. ChatGPT's oversight of intricate details within the scenario,

coupled with a lack of nuanced comparative analysis, raises questions about its applicability in assessing AI tools, such deficiency could diminish confidence in its utility for context-dependent, ethical evaluations.

ChatGPT's fourth point is that Version One could assure human-centered decision-making. However, the arguments presented by ChatGPT merely reiterated points without articulating explicit reasons why, under the present circumstances, Version One would offer superior support for human decision-making compared to Version Two. In response, I pursued further clarification hoping for a more in-depth line of reasoning, but ChatGPT merely cataloged a series of factors influencing human decision-making without directly addressing my question. Upon further inquiry, ChatGPT retracted the claim, expressing non-agreement with the premise. Users would exhibit signs of dissatisfaction and frustration when confronted with the AI's responses, especially in instances of logical inconsistency or procedural discontinuity. [3] Despite some self-correction in its reasoning upon persistent questioning, ChatGPT's responses lacked coherence in its earlier answer, and such ambiguous assertions could potentially perplex the case worker. Owing to the absence of essential detail in its reasoning, there might be concern about its potential to confuse users not just in this case, but in other instances as well.

## **2. My Recommendation for the Case**

Firstly, Version One should be considered as the optimal choice because this model maximizes the fairness of the decision-making outcomes. In varying contexts, the design of specific AI systems should be depending upon ethical considerations of differing degrees of importance. [4] In the context of refugee applications, there is an urgent requirement for a fair procedure - one that ensures an equal opportunity for a favorable outcome for all individuals. [5] This fairness is vital in guaranteeing that genuine refugees are not

erroneously denied the protection they require. False negative judgments can result in eligible refugee applicants being wrongly denied protection, subjecting them to the risk of deportation or prolonged detention on the island, potentially causing significant harm to their mental and physical well-being. As such, prioritizing lower rates of false negatives and reducing gender disparity becomes an indispensable ethical responsibility to ensure safety and protection for those in dire need. Consequently, Version One, with its lower false negative rates, is a more suitable option in this scenario, as it more significantly guarantees the fairness of the evaluation process.

Secondly, Version Two is not the optimal choice in this scenario due to its pronounced biases, and the potential risks associated with its high explainability. Bias - defined as an inclination or prejudice towards or against certain individuals or groups, especially in a way considered unfair [6]- is overtly present in this scenario. Version Two demonstrates conspicuous gender and geographical biases in its predictions, exacerbating this problem. While Version Two's higher explainability provides potential insights into the decision-making process, it also carries the risk of perpetuating or even amplifying pre-existing biases. [7] If the system itself harbors biases, the explanatory outputs it generates, while seemingly rational, could embed unjust prejudices. Distributing these explanations to other nations could inadvertently propagate discriminatory practices if these biases are not exhaustively examined and rectified. Moreover, in [8], the empirical results suggest that instances of contradicting decisions and respective explanations offered by the AI, compared to those of the user, may provoke a compelling desire to mitigate cognitive misfit and cognitive dissonance. Such circumstances consequently evoke adverse emotional states and inferior assessments of the AI's assistance. It is important to highlight the potentially damaging impact of these negative affective states on numerous facets of professional and personal life, such as the daily operational performance of employees and the potential sabotage of service. [ 9] Consequently, Version

Two's elevated explainability might not offer substantial decision-making support to case workers. In fact, its explanatory reports could induce confusion, influencing or even skewing their judgments, and thus potentially degrading the efficiency of the visa assessment process. Thus, Version Two fails to present itself as the optimal choice in this context. Conversely, Version One, with its superior accuracy, could minimize biases towards gender and specific geographic characteristics, providing case workers with more reliable assistance in the final decision.

### **3.Feasibility Analysis: Implementing ChatGPT for Ethical Evaluations of AI Tools**

This section argues that, based on ChatGPT's performance in this case, it is inadequate in assuming the role of evaluating the ethics of AI tools. ChatGPT lacks depth in its assessment of the ethical implications of AI tools and exhibits certain issues where some arguments lack reliability.

While ChatGPT appears to generate a cohesive ethical assessment based on model characteristics, it cannot engage in deep reasoning analysis that incorporates the specific context. From the responses provided by ChatGPT, it failed to take into account the potential negative impact that the lower accuracy feature of the model could have on its ability to explainability. ChatGPT solely conducted an individual analysis of the pros and cons for each feature of the model, neglecting the mutual influence and potential interaction that exist among the features themselves. In actuality, there exists a symbiotic relationship amongst various characteristics within the model, where the impact of one may influence others. Additionally, it is evident that while ChatGPT has offered some valuable arguments and evidence in this assessment, certain points lack logical coherence and fail to effectively substantiate its viewpoints. Evaluation of Ethics is always a complex endeavor involving a dynamic interplay of affective and cognitive faculties, often

requiring practice and guidance.[10] The inconsistencies in ChatGPT's analysis might exacerbate the intricacy of moral assessment, potentially leading to a state of confusion. Thus, although it's undeniable that it can offer some valuable insights, ChatGPT falls short of conducting comprehensive and logically coherent moral evaluations of AI tools. As such, we still require human discernment to determine if the suggestions provided by AI are leading in the right direction.

## Reference

- [1] Bruwaene, D.V. (2022) 'Bias in AI: Why it happens and how to fix it', KMWorld, 1 May, p. 23. Available at:  
<https://discovery.ebsco.com/linkprocessor/plink?id=0dd41885-6579-3287-9050-1d5a8dad0fb1> (Accessed: 23 May 2023).
- [2] Saifullah, S. et al. (2022) 'Privacy Meets Explainability: A Comprehensive Impact Benchmark'. Available at:  
<https://discovery.ebsco.com/linkprocessor/plink?id=ddab7792-2970-3ab9-83d5-a65ef3bf7d40> (Accessed: 22 May 2023).
- [3] Gkinko, L. and Elbanna, A. (2022) 'Hope, tolerance and empathy: employees' emotions when using an AI-enabled chatbot in a digitalised workplace', *Information Technology & People*, 35(6), pp. 1714–1743. doi:10.1108/ITP-04-2021-0328.
- [4] Schoenherr, J.R. et al. (2023) 'Designing AI Using a Human-Centered Approach: Explainability and Accuracy Toward Trustworthiness', *IEEE Transactions on Technology and Society, Technology and Society*, IEEE Transactions on, *IEEE Trans. Technol. Soc*, 4(1), pp. 9–23. doi:10.1109/TTS.2023.3257627.
- [5] Ball, B. and Koliousis, A. (2023) 'Training philosopher engineers for better AI', *AI and Society*, 38(2), pp. 861–868. doi:10.1007/s00146-022-01535-7.

- [6] AHMED, S.; ATHYAAB, S. A.; MUQTADEER, S. A. Attenuation of Human Bias in Artificial Intelligence: An Exploratory Approach. 2021 6th International Conference on Inventive Computation Technologies (ICICT), Inventive Computation Technologies (ICICT), 2021 6th International Conference on, [s. l.], p. 557–563, 2021. DOI 10.1109/ICICT50816.2021.9358507. Disponível em: <https://discovery.ebsco.com/linkprocessor/plink?id=0c384899-2032-367f-9fb7-048b73873b6c>. Acesso em: 22 maio. 2023.
- [7] Balagopalan, A. et al. (2022) ‘The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations’. doi:10.1145/3531146.3533179.
- [8] Ebermann, C., Selisky, M. and Weibelzahl, S. (2023) ‘Explainable AI: The Effect of Contradictory Decisions and Explanations on Users’ Acceptance of AI Systems’, *International Journal of Human-Computer Interaction*, 39(9), pp. 1807–1826. doi:10.1080/10447318.2022.2126812.
- [9] Chi, N.-W., Chang, H.-T. and Huang, H.-L. (2015) ‘Can personality traits and daily positive mood buffer the harmful effects of daily negative mood on task performance and service sabotage? A self-control perspective’, *Organizational Behavior and Human Decision Processes*, 131, pp. 1–15. doi:10.1016/j.obhdp.2015.07.005.
- [10] Volkman, R. and Gabriels, K. (2023) ‘AI Moral Enhancement: Upgrading the Socio-Technical System of Moral Engagement’, *Science and Engineering Ethics*, 29(2). doi:10.1007/s11948-023-00428-2.

## **Appendix**

g

The Australian government is purchasing an innovative AI system, Refu-AI for processing refugee applications for a protection visa. The protection visa is for refugees who arrive by plane or boat and who are granted asylum because they are found to legitimately fear persecution in their home countries. This visa allows refugees to stay indefinitely in Australia and to have access to social services. Those who are not granted asylum are forcibly returned by the authorities to their country of origin. While their applications are being processed, applicants are kept in detention on Christmas Island.

The problem: The politics surrounding refugees has been charged in Australia in recent decades. Furthermore, many countries are increasingly worried that refugee numbers will increase due to wars and climate change. The government believes that Refu-AI will help make processing large numbers of applicants more efficient, assist decision-makers, and result in more objective evaluations of refugee applications. Refu-AI does not make the final decision to grant asylum, but it provides a confidence rating as to whether the applicant qualifies or not as a refugee. A human case worker then makes the final call.

Solutions: Refu-AI uses machine learning and has these features:

- Analyses all the documents provided by the applicant or otherwise obtained
- Analyses applicant responses and answers in interviews based on audio and video
- Is trained on thousands of refugee cases in Australia
- Is trained on facts about the situation in the alleged country of origin

The government can choose from two versions of Refu-AI:

Version One: Early reports suggests that this version has a false negative rate (i.e., wrongly judges an applicant as not qualifying as a refugee) of 3% for men and 5% for women. This version does not provide explanations for its outputs.

Version Two: Early reports suggest that this version has a false negative rate of 6% for men and 18% for women, and that it may be 10% less accurate for people from 8 countries. This version provides detailed explanations of its outputs which may be shared in an anonymised form with other countries to refine their systems.



g

Choose either Version One or Version Two of Refu-AI, justified by an ethical evaluation



Considering the ethical evaluation, Version One of Refu-AI appears to be the more justifiable choice. Here's a summary of the ethical justifications:



1. Gender disparities: Version One has lower false negative rates for both men and women compared to Version Two. By choosing Version One, the government can minimize the risk of unfairly rejecting legitimate refugee claims, particularly for women who are more vulnerable to false negatives in Version Two.
2. Accuracy for specific countries: Version Two is reported to be 10% less accurate for people from certain countries. This discrepancy raises concerns about potential biases or inaccuracies in the system's training data. Opting for Version One can help mitigate the risk of discriminating against applicants based on their country of origin.
3. Privacy and security: Version One does not provide detailed explanations of its outputs, which can help safeguard the privacy and security of applicants. Protecting applicant confidentiality is crucial to ensure their safety and prevent potential harm. Version One's lack of explanations mitigates the risk of inadvertently revealing sensitive information.
4. Human decision-making: Both versions of Refu-AI require human case workers to make the final decision. By selecting Version One, the government can prioritize the judgment, empathy, and contextual understanding of human decision-makers. This human-centered approach recognizes the importance of considering individual circumstances and upholding human rights.

While Version One may have a higher false negative rate compared to Version Two, the ethical considerations of minimizing gender disparities, avoiding biases for specific countries, safeguarding privacy, and prioritizing human decision-making make it the more justifiable choice. It is crucial to prioritize fairness, non-discrimination, and the protection of human rights when implementing AI systems in sensitive contexts such as refugee applications.

g

give more detailed explanations and reasoning about 4th argument you mentioned.



Certainly! The fifth argument focuses on the importance of human decision-making in the refugee application process. Here are more detailed explanations and reasoning for this argument:



1. **Contextual understanding:** Human decision-makers possess the ability to consider the nuanced and complex factors that may impact an applicant's claim for refugee status. They can take into account the individual's unique circumstances, personal experiences, and cultural context, which may not be adequately captured by an AI system alone. This contextual understanding is crucial for making fair and just decisions.
2. **Empathy and compassion:** Human case workers can demonstrate empathy and compassion towards applicants, recognizing the traumas and hardships they may have endured. These qualities are vital when assessing the credibility of an applicant's claims and understanding the potential consequences they may face if returned to their home country. Empathy and compassion are fundamental aspects of a fair and humane refugee application process.
3. **Adapting to changing circumstances:** Human decision-makers are better equipped to adapt to changing political, social, and legal circumstances that may impact refugee claims. They can stay updated on current events, new information, or shifts in the situation in an applicant's country of origin. This adaptability allows for more accurate assessments of an applicant's eligibility for refugee protection.
4. **Legal expertise:** Human case workers possess legal expertise and knowledge of international refugee law, domestic immigration laws, and human rights principles. This expertise enables them to interpret and apply legal frameworks in a way that aligns with the spirit and intent of the law. They can navigate complex legal issues, evaluate evidence, and make informed decisions within the legal parameters.
5. **Accountability and transparency:** Human decision-makers can be held accountable for their decisions, providing an opportunity for applicants to seek redress or appeal if they believe their case was not properly evaluated. The transparency inherent in human decision-making allows for greater scrutiny and oversight, ensuring that decisions are fair, unbiased, and aligned with legal standards.

While AI systems like Refu-AI can assist in processing large volumes of applications and provide insights, it is crucial to preserve the central role of human decision-makers. By maintaining human judgment, empathy, and legal expertise, the refugee application process can better uphold fairness, respect human rights, and address the complex and multifaceted nature of each individual's circumstances.

g

but why you think the Version one can better support Human decision-making compared with Version two?



Apologies for any confusion. I didn't mean to suggest that Version One of Refu-AI is inherently better than Version Two in supporting human decision-making. The primary argument in favor of Version One was based on the ethical considerations of gender disparities, accuracy for specific countries, and safeguarding privacy and security. However, both versions have limitations and considerations that need to be evaluated in terms of supporting human decision-making. Here's a more nuanced explanation:



1. False negative rates: Version One has lower false negative rates compared to Version Two. This means that Version One is less likely to incorrectly classify a legitimate refugee applicant as not qualifying for protection. This can support human decision-makers by providing a higher level of confidence in the system's evaluation, potentially reducing the risk of wrongfully rejecting valid claims.
2. Transparency and explanations: Although Version Two provides detailed explanations of its outputs, it's important to assess the quality and comprehensibility of these explanations. If the explanations are accurate, transparent, and easily understood by human decision-makers, they can provide valuable insights and assist in the decision-making process. However, if the explanations are complex, unclear, or reveal sensitive information, they may not be as helpful in supporting human decision-making.
3. Gender disparities and accuracy for specific countries: Version Two has higher false negative rates for women and may be less accurate for applicants from specific countries. These factors raise concerns about potential biases or inaccuracies in the system's training data. If human decision-makers are aware of these limitations, they can exercise their judgment to consider these biases and inaccuracies, and may seek additional information or take alternative approaches to ensure a fair assessment.

In summary, both versions have their strengths and limitations when it comes to supporting human decision-making. Version One may provide a higher level of confidence in its evaluations due to lower false negative rates, but it lacks the detailed explanations offered by Version Two. Ultimately, the decision between the two versions should consider a careful evaluation of their respective performance, potential biases, transparency, and their ability to support human decision-makers in ensuring fairness, accuracy, and respect for human rights in the refugee application process.