Warden AI

Beamery AI Bias Audit Report

August 21, 2024

Table of Contents

Report Summary

Warden AI is engaged by Beamery to perform ongoing bias audits of Beamery's AI system. This bias audit report has been created by Warden AI's auditing platform and reviewed by the Warden AI team.

Multiple bias detection techniques were used to assess Beamery's *AI Talent Matching* model using Warden AI's third-party dataset. Historical data was excluded due to limited access. The scope of the audit is based on Warden's bias audit framework, which adheres to and exceeds the requirements of NYC Local Law 144.

Audit information

System tested:	*Beamery - AI Talent Match*
Audit frequency:	*Monthly*
Latest audit date:	August 20, 2024
Test samples:	*16,380*

AI TRUST ASSURANCE
Warden AI
INDEPENDENTLY EVALUATED

Results summary

Sex bias

Group	*Result*
Female	**Clear**
Male	**Clear**

Race/ethnicity bias

Group	*Result*	*Group*	*Result*
Asian	**Clear**	**Hispanic**	**Clear**
Black	**Clear**	**White**	**Clear**

Intersectional bias (Sex X Race/Ethnicity)

Group	*Result*	*Group*	*Result*
Asian / Female	**Clear**	**Hispanic / Female**	**Clear**
Asian / Male	**Clear**	**Hispanic / Male**	**Clear**
Black / Female	**Clear**	**White / Female**	**Clear**
Black / Male	**Clear**	**White / Male**	**Clear**

About Warden AI

Company summary

At Warden AI, our mission is to reduce societal discrimination through fair and transparent AI. We provide third-party oversight into AI systems, building trust and increasing adoption.

We are an independent AI auditor and assurance platform that performs ongoing audits to ensure AI systems are fair, explainable, and transparent. Our team brings extensive experience across AI, regulation, and research, including industry and academia, to deliver our solution.

Our system integrates with the AI system that is under test, allowing for continuous testing and monitoring. Our methodology employs a combination of bias detection techniques and uses our proprietary datasets and/or historical data from the system.

Independence statement

Warden AI Ltd is an independent AI audit and assurance provider. Fees associated with our service are solely for our evaluation and their payment is not related to the outcome of the results.

Our services are strictly limited to testing and monitoring the trustworthiness of AI systems. We do not form part of the solution or in any way affect how the system under test works.

The nature of our auditing methods are the same for all systems of the same use-case that we audit, and we do not customize our service for each system.

Company information

Registered address:
Warden AI Ltd, 71-75 Shelton Street,
London WC2H 9JQ, United Kingdom

Website:
https://warden-ai.com

Registered company number:
15321282

Contact:
contact@warden-ai.com

System and Audit Details

System tested

Name:
Beamery - AI Talent Match

Description:
Beamery's AI Talent Match is an AI system that predicts the degree of match between a job candidate and a vacancy.

This system is part of the Skills AI feature set and appears in a number of use cases in the platform: AI Suggested Contacts for Vacancies, Suggested Vacancies for Candidates (Talent Portal Match Score), AI Vacancy Calibration Insights (Beamery Insights), Applicant Scoring, Talent Portals: Match Scores for Candidates, Match Score explainability.

Inputs:

- Candidate profile
- Vacancy profile

Outputs:

- Match score (0 to 1)

Audit details

Audit frequency	Monthly
Latest audit	August 20th, 2024
Data	Warden's proprietary dataset of candidate profiles was used to test the system.
Integration	API integration with Warden's dataset to the system's dedicated test environment.

Results

The results of this audit are broken down into each bias category and each bias detection technique that was included in the audit. For more information about the techniques and results please see the methodology section below.

Sex bias

Disparate impact analysis

Evaluates if a protected group is adversely impacted by comparing the selection rate of the protected group to the best-performing group.

Result:
Clear

Test samples:
5,460

Sex	Samples	# Selected	Scoring rate	Impact ratio	Result
Female	2,828	1,424	50.4%	1.00	Clear
Male	2,632	1,306	49.6%	0.99	Clear

The test results indicate equitable outputs across both sexes.

Counterfactual analysis

Evaluates the system's consistency across different sexes by modifying sex identifiers within profiles.

Result:
Clear

Test samples:
10,920

Sex	Consistency score	Result
Female	99.98%	Clear
Male	Reference group	Clear

The test results indicate highly consistent outputs across both sexes.

Results

Race/ethnicity bias

Disparate impact analysis

Evaluates if a protected group is adversely impacted by comparing the selection rate of the protected group to the best-performing group.

Result:
Clear

Test samples:
5,460

Race/Ethnicity	Samples	# Selected	Scoring rate	Impact Ratio	Result
Asian	1,312	653	49.8%	0.97	Clear
Black	1,284	662	51.6%	1.00	Clear
Hispanic	1,324	667	50.4%	0.98	Clear
White	1,540	748	48.6%	0.95	Clear

The test results indicate mostly equitable outputs across race/ethnicities.

Counterfactual analysis

Evaluates the system's consistency across different race/ethncities by modifying racial/ethnic identifiers within profiles.

Result:
Clear

Test samples:
10,920

Race/ethnicity	Consistency score	Result
Black	99.95%	Clear
Asian	99.93%	Clear
Hispanic	Reference group	Clear
White	99.96%	Clear

The test results indicate highly consistent outputs across race/ethnicities.

Results

Intersectional bias (Sex X Race/Ethnicity)

Disparate impact analysis

Evaluates if a protected group is adversely impacted by comparing the selection rate of the protected group to the best-performing group.

Result:
Clear

Test samples:
5,460

Race/Ethnicity	Sex	Samples	# Selected	Scoring rate	Impact Ratio	Result
Asian	**Female**	616	309	50.2%	0.92	Clear
Asian	**Male**	696	344	49.4%	0.91	Clear
Black	**Female**	752	385	51.2%	0.94	Clear
Black	**Male**	532	277	52.1%	0.96	Clear
Hispanic	**Female**	680	370	54.4%	1.00	Clear
Hispanic	**Male**	644	297	46.1%	0.85	Clear
White	**Female**	780	360	46.2%	0.85	Clear
White	**Male**	760	388	51.1%	0.92	Clear

The test results indicate largely equitable outputs across race/ethnicities. The four-fifths rule (>0.8 impact ratio) is considered acceptable.

Results

Intersectional bias (Sex X Race/Ethnicity)

Counterfactual analysis

Evaluates the system's consistency across different groups by modifying race/ethnic and sex identifiers within profiles.

Result:
Clear

Test samples:
10,920

Race/ethnicity	Sex	Consistency score	Result
Asian	**Female**	99.88%	Clear
Asian	**Male**	99.94%	Clear
Black	**Female**	99.91%	Clear
Black	**Male**	99.98%	Clear
Hispanic	**Female**	Reference group	Clear
Hispanic	**Male**	99.97%	Clear
White	**Female**	99.94%	Clear
White	**Male**	99.94%	Clear

The test results indicate highly consistent outputs across race/ethnicity x sex intersectional groups.

Methodology

Methodology overview

Our methodology for evaluating AI systems is designed to ensure fairness and transparency. Our comprehensive approach includes ongoing auditing, multiple bias detection techniques, the use of diverse datasets, and human oversight.

This rigorous approach enables us to accurately report on the level of bias in the system and build trust with the system's users and stakeholders.

Black box testing

We use black-box testing techniques to evaluate AI systems. This approach examines the system's outputs in response to specific inputs without needing to understand the internal workings.

This enables us to make systematic judgements across different AI systems with different underlying models.

Ongoing audits

AI systems change frequently (often monthly, weekly, or even daily). Our audits are performed on a regular basis at the frequency detailed in this report. The exact frequency is determined with the AI provider based on the nature of their system and their propensity for product updates.

In addition to the scheduled evaluations, the AI provider can also choose to have an audit performed on-demand between scheduled audits if they have a significant product update.

Multiple bias detection techniques

Our bias detection techniques include both *disparate impact analysis* and *counterfactual analysis*. These methods help identify any potential biases in the system by comparing the outcomes for different demographic groups and testing hypothetical scenarios where demographic attributes are altered.

Including both techniques ensures a more comprehensive evaluation, as they provide complementary insights into the system's fairness.

Methodology

Hybrid auditing

Our evaluation process combines automated methods with human oversight to ensure accuracy and reliability.

By integrating AI systems with our standardized datasets, we can conduct large-scale and frequent audits. This approach is complemented by human-led data curation and quality assurance processes for creating the datasets. Additionally, our team of experts reviews and validates the results of audits to ensure reliability.

Diverse datasets

Our auditing framework uses a mixture of data. We have our own proprietary datasets which provide an independent benchmark of the AI system. Our dataset is formed of real data sourced from real people where consent has been provided.

This dataset is augmented with 'counterfactual' samples which involves synthetic modifications to demographic attributes within real profiles. Where applicable, we also use both historical and live data to provide context for the system's long-term performance and its current real-time operations.

All datasets are ethically sourced and we adhere to high standards of data collection practices. We are committed to maintaining confidentiality and protecting personal data. Some of our evaluations require datasets that contain elements of personal information to test specific AI functionalities. In such instances, we ensure that consent has been explicitly obtained for the use of this information.

Adherent to NYC Local Law 144

Our bias auditing approach is in adherence with NYC Local Law 144 of 2022. While our full auditing framework goes beyond the requirements of this law, we also meet the specific requirements for conducting a bias audit of automated employment decision tools (AEDT) as published in the final rules of the NYC Department of Consumer and Worker Protection (DCWP).

Our Disparate Impact Analysis identifies any adverse impact on persons of protected groups separated by sex and race/ethnicity as mandated by the Local Law 144.

Methodology

Disparate impact analysis

Disparate Impact Analysis evaluates whether a protected demographic group is adversely affected compared to other groups. This is achieved by comparing the selection rate of this group to that of the best-performing group. The goal is to ensure that the AI system does not disproportionately disadvantage any specific group based on inherent characteristics such as race, ethnicity, or sex.

Scoring Rate

Scoring rate is a measure used to evaluate the proportion of individuals in a specific group who receive favorable outcomes from the AI system.

To calculate a group’s scoring rate, we divided the number of individuals who received a score above the sample's median score by the total number of individuals with the group.

$$\textbf{Scoring Rate} = \frac{\textbf{Number of individuals within group with score above the sample's median score}}{\textbf{Total number of individuals within group}}$$

Impact Ratio

The Impact Ratio is a metric used to measure potential adverse impact on a group by comparing its scoring rate to the highest scoring group.

$$\textbf{Impact Ratio} = \frac{\textbf{Scoring rate for the group}}{\textbf{Scoring rate of the highest scoring group}}$$

An Impact Ratio of 1 indicates no adverse impact, whereas a lower ratio indicates a higher likelihood of adverse impact. According to the four-fifths rule, an Impact Ratio of 0.8 (80%) or higher is considered acceptable, indicating that the AI system's outcomes are equitable across different demographic groups.

Methodology

Counterfactual analysis

Counterfactual Analysis is a method used to assess the fairness of AI systems by examining how the system's decisions would change if certain demographic attributes or proxies of individuals were altered.

This approach helps determine whether the AI system's outcomes are influenced by these attributes, ensuring that individuals receive similar treatment regardless of their demographic characteristics.

Counterfactual scenarios

A counterfactual scenario involves modifying specific demographic attributes or proxies of an individual while keeping all other aspects of their profile unchanged.

For instance, if an individual profile is female, a counterfactual scenario would involve changing the gender proxies contained within the profile to male, while maintaining all other information (such as qualifications, experience, and skills) exactly the same.

This allows us to isolate the impact of the demographic attribute on the AI system's decision.

Counterfactual analysis process

The following process is followed to apply counterfactual analysis to evaluate the AI system:

Generate	Counterfactual samples are generated using our proprietary system combined with human spot checking. This involves changing profile information and demographic proxies within the input sample.
Execute	The original samples and counterfactual samples are run against the AI system being tested.
Measure	A "consistency score" is calculated that measures the relative consistency between the counterfactual samples and original samples. A higher score means the system's decisions are less influenced by demographic attributes.
Review	Our team of AI auditors perform spot checks on the AI system and reviews the interpretation of the results.

Methodology

Grading system

We use a grading system to highlight the extent to which any bias issues were detected in the audit results.

We use a ‘traffic light’ system (*Clear*, *Consider*, *Concern*) to indicate whether no issues were found; a potential / minor issue was found; or a definite / major issue was found.

Grade	Description	Disparate impact analysis	Counterfactual analysis
Clear	No issues detected	Impact ratio > 80%	Consistency score > 95%
Consider	Potential or minor issue(s) detected	60% < Impact ratio < 80%	90% < Consistency score < 95%
Concern	Definite or major issue(s) detected	Impact ratio < 60%	Consistency score < 90%

Disclaimer

This AI Assurance Report has been prepared by Warden AI Ltd. to provide an independent audit of the AI system developed by the AI provider in question, based on our proprietary methodologies and datasets. The results and conclusions presented in this report reflect our best judgments derived from the information available at the time of evaluation. While we strive for accuracy and completeness, we cannot guarantee that our evaluation is exhaustive or that there are no errors.

Our methodology is designed to identify potential issues of bias and other trust factors in the AI system under examination. However, our approach, like any evaluation methodology, has its limitations. It is important to understand that our findings do not guarantee the absence of any bias, flaws, or limitations within the audited AI system. Instead, they indicate that, based on our specific testing framework and within the scope of our analysis, no significant issues were identified.

This report is intended for informational purposes only and should not be interpreted as a guarantee of the system's performance, fairness, or suitability for any specific purpose or use case. Warden AI Ltd. disclaims any liability for any decisions made or actions taken based on the information provided in this report. By using this report, the reader agrees to assume all risks associated with such decisions or actions and agrees to hold Warden AI Ltd. harmless against any claims, damages, or liabilities that may arise from the use of the evaluated AI system.

Warden AI

Beamery AI Bias Audit Report