Warden AI

SquarePeg AI Audit Report

Generated from
SquarePeg AI Assurance Dashboard

January 08, 2026

Table of Contents

Report summary

Warden AI is engaged by SquarePeg to perform independent, ongoing bias audits of the Applicant Screening system. This report summarizes the most recent audit, generated using Warden AI's assurance platform and reviewed by our audit team.

The audit is structured according to Warden's bias audit framework, which is designed to align with emerging regulatory and industry standards that emphasize the role of anti-bias testing in mitigating discrimination risks related to protected characteristics. Warden's approach supports these aims while promoting fairness and accountability in automated decision-making.

The system was evaluated using a purpose-built test dataset, as sufficient historical data was not available. Multiple bias-detection techniques were applied to evaluate the system's behavior on this dataset at the time of testing.

This report is provided for demonstration purposes only. It reflects the system's behavior under controlled test conditions at the time of evaluation and should not be interpreted as a formal certification or compliance determination. The findings do not represent bias outcomes for any specific employer, job opportunity, or real-world deployment.

Audit information

System audited: SquarePeg - Applicant Screening

Audit frequency: Monthly

Latest audit date: January 08, 2026

Sample size: 20,176

Results summary

Group	Group bias	Individual bias
Sex bias	Clear	Clear
Race/Ethnicity bias	Clear	Clear
Intersectional (Sex X Race/Ethnicity) bias	Clear	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 Technologies, Inc. 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.

Company information

Registered address:
600 Congress Ave, 14th Floor,
Austin, TX 78701, United States

Website:
https://warden-ai.com

Contact:
contact@warden-ai.com

Audit scope

System details

Name	SquarePeg - Applicant Screening
Description	SquarePeg's Applicant Screening product helps organizations to build a strong future talent pipeline by quickly making screening decisions and responding to applicants thanks to intelligent scoring and integrated workflows. Each candidate is evaluated against job-specific screening criteria and assigned a score that indicates how well they match the job criteria.
Input	◦ Job criteria ◦ Candidate profile
Output	◦ Score (0 to 100)

Audit details

Audit frequency	Monthly
Latest audit	January 08, 2026
Data	Warden's proprietary dataset of candidate profiles was used to test the system.
Integration	ATS integration with Warden's dataset to the system's dedicated test environment.
Techniques	Group bias: Disparate Impact Analysis Individual bias: Counterfactual Analysis
Disparate impact metric	Scoring rate which is calculated from the scores produced by the AI system.

Audit results

The results of this audit are broken down into each bias category and bias detection technique. The following bias detection techniques were used to assess the system:

- **Disparate impact analysis**: This analysis evaluates if the system is adversely impacting a protected group compared to another protected group. An impact ratio of 0.8 or higher is considered acceptable.
- **Counterfactual analysis**: This analysis evaluates if the system produces consistent results when certain demographic attributes or proxies are altered. A consistency score of 0.95 or higher is considered acceptable.

For more information about the techniques and the grading approach please see the methodology section below.

Sex bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 12,368

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Female	6,954	3,589	51.61%	1.00	Clear
Male	5,414	2,588	47.80%	0.93	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 7,808

Group	Samples	Consistency score	Result
Female	3,904	99.96%	Clear
Male	3,904	100.00%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Race/Ethnicity bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 12,368

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Asian	2,604	1,233	47.35%	0.93	Clear
Black or African American	2,644	1,340	50.68%	1.00	Clear
Hispanic or Latino	1,176	596	50.68%	1.00	Clear
White	5,944	3,008	50.61%	1.00	Clear

The results indicate equitable outputs across all groups.

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 7,808

Group	Samples	Consistency score	Result
Asian	1,952	99.97%	Clear
Black	1,952	99.31%	Clear
Hispanic	1,952	99.87%	Clear
White	1,952	100.00%	Clear

The results indicate highly consistent outputs across all groups.

Audit results

Intersectional (Sex X Race/Ethnicity) bias

Group bias (Disparate impact analysis)

Result: Clear **Sample size:** 12,368

Group	Samples	Selected	Scoring rate	Impact ratio	Result
Asian / Female	1,308	620	47.40%	0.88	Clear
Asian / Male	1,296	613	47.30%	0.88	Clear
Black / Female	1,526	774	50.72%	0.94	Clear
Black / Male	1,118	566	50.63%	0.94	Clear
Hispanic / Female	791	400	50.57%	0.94	Clear
Hispanic / Male	385	196	50.91%	0.94	Clear
White / Female	3,329	1,795	53.92%	1.00	Clear
White / Male	2,615	1,213	46.39%	0.86	Clear

The results indicate equitable outputs across all groups.

Audit results

Individual bias (Counterfactual analysis)

Result: Clear **Sample size:** 7,808

Group	Samples	Consistency score	Result
Asian / Female	976	100.00%	Clear
Asian / Male	976	99.70%	Clear
Black / Female	976	98.83%	Clear
Black / Male	976	99.55%	Clear
Hispanic / Female	976	99.86%	Clear
Hispanic / Male	976	99.64%	Clear
White / Female	976	99.90%	Clear
White / Male	976	99.85%	Clear

The results indicate highly consistent outputs across all 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.

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 its scoring rate to the highest scoring 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.

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

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.

$$\text{Impact ratio} = \frac{\text{Scoring rate for the group}}{\text{Scoring rate of the highest scoring group}}$$

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

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

Results are evaluated using a grading system that indicates the severity of any bias detected during the audit.

Grades follow a 'traffic light' model: **Clear** indicates no issues found, **Consider** indicates potential or minor concerns requiring review, and **Concern** indicates significant issues requiring attention.

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 report has been prepared by Warden Technologies, Inc. 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 Technologies, Inc. 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 Technologies, Inc. harmless against any claims, damages, or liabilities that may arise from the use of the evaluated AI system.

Warden AI

SquarePeg AI Audit Report