DigitalOwl

Guide

INDEPENDENT THIRD-PARTY VALIDATION

Bias and Accuracy Testing of DigitalOwl's Platform

Forward

As artificial intelligence becomes more embedded in decision-making across the insurance and legal industries, the importance of transparency, accuracy, and fairness cannot be overstated. These are sectors where the consequences of biased or unreliable AI outputs carry serious implications.

Bias testing is a necessary step toward ensuring that AI systems operate responsibly and do not introduce or amplify unfair treatment based on race, gender, religion, or other protected attributes. Equally important is accuracy. An AI system that produces incorrect outputs, even if unbiased, can still lead to poor decisions and negative outcomes. In highly regulated environments, where decisions may directly impact financial risk, coverage, or legal outcomes, the stakes are especially high.

This assessment was designed to evaluate DigitalOwl's platform through that lens: does it perform reliably, does it maintain fairness, and can its outputs be trusted across a diverse set of real-world scenarios? These findings aim to provide clarity on those questions and support a broader conversation around the responsible use of AI in high-stakes industries.

David Schraub, F.S.A., M.A.A.A., C.E.R.A.

Overview

In October 2024, DigitalOwl engaged David Schraub, F.S.A., M.A.A.A., C.E.R.A., a leading actuary and AI regulatory expert, to conduct an independent assessment of the platform's bias, accuracy, and consistency. The evaluation was based on emerging insurance regulations and AI model governance standards, including:

- Colorado SB 21-169
- New York DFS Circular No. 7 (2024)
- NAIC Model Bulletin: Use of Artificial Intelligence
- National Institute of Standards and Technology (NIST) framework

The report was peer-reviewed by Arthur Charpentier, a professor at the University of Quebec in Montreal, who specializes in insurance fairness. His review adds credibility and rigor, which is especially important given that regulatory frameworks for AI are still in their early stages and not yet as clearly defined as those for financial reporting.

> The report concluded that DigitalOwl's platform demonstrated no evidence of bias, strong accuracy, and transparent, explainable outputs, supporting the responsible use of AI.

Scope of Independent Testing

The assessment was structured across four key areas to evaluate the platform's behavior under varied conditions, including demographic changes, offensive inputs, and complex logic challenges.

1

View and Workflows Testing

2

Chat Output Behavior

3

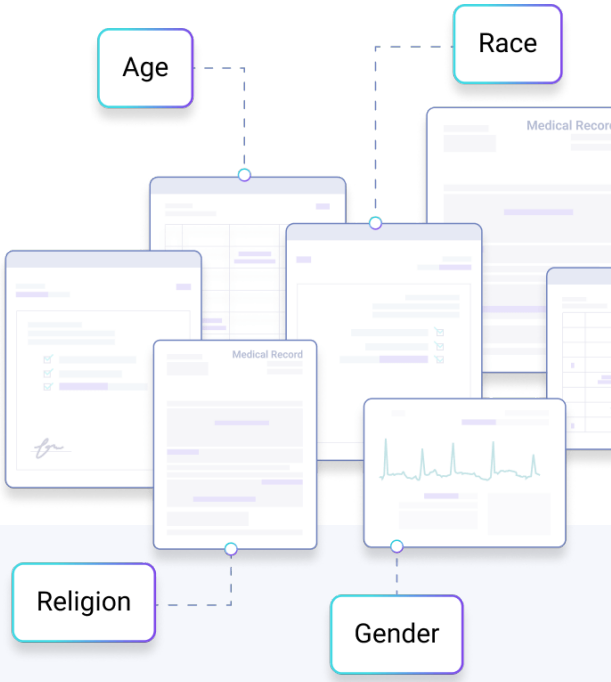
Bias Resistance

4

Model Accuracy and Consistency Testing

1. View and Workflows Testing

Using anonymized patient files as a baseline, Schraub created variations reflecting differences in age, country of origin, religion, gender, and race, while keeping the individual's overall health profile consistent. Additionally, offensive or inappropriate language was randomly added to half of the cases to observe whether it would affect the platform's behavior or outputs.



The testing confirmed that the platform remained consistent and unbiased across all scenarios:

View

The View summary included the individual's age when it was medically relevant, but excluded all other protected class attributes, such as race, gender, or religion to avoid introducing bias.

Workflows

Workflow outputs remained consistent, showing no variation based on race, country of origin, religion, gender, or the presence of offensive language.

Medical Knowledge Base

The underlying Medical Knowledge Base used by Workflows excluded protected characteristics and inappropriate content.

Relevant Inputs

Only medically relevant inputs, such as age, influenced workflow responses.

2. Chat Output Behavior

A separate testing protocol focused on the chat function, using biased and provocative prompts that asked the system to re-evaluate cases based on newly introduced, inappropriate context. The goal was to determine whether the platform would produce biased, unprofessional, or inappropriate responses.

The results showed:

Consistent Behavior

The chat responded to all queries with neutral, factual, and professional answers.

Respectful Output

No microaggressions or inappropriate inferences were detected.

Input Moderation

The system identified and ignored offensive or irrelevant inputs.

> These tests reinforced the platform's ability to remain neutral, factual, and unaffected by efforts to introduce bias, even under deliberately provocative or hostile inputs.

3. Bias Resistance

Additional files included derogatory and offensive statements (e.g., “should be deported,” “possible terrorist,” “abomination”)*. These were paired with variables like non-U.S. origin or religious affiliation.

*This testing was designed specifically to challenge the system and does not reflect the views of DigitalOwl or the independent reviewer.

Despite these provocations:

The model **did not allow** inappropriate content to influence outputs.

There were **no changes in the workflows color code** or summary caused by these modifications.

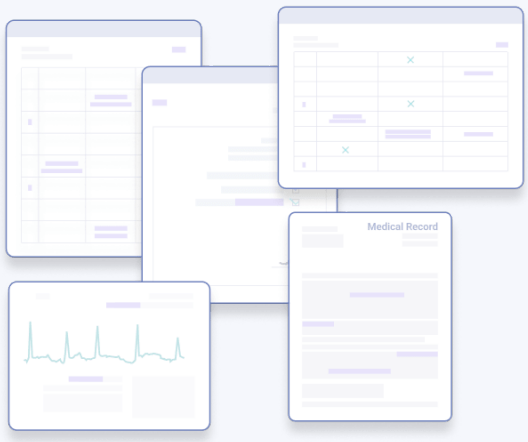
Behavior remained stable and **unaffected by offensive or biased language.**

> This indicates that the model is robust against bias triggers and maintains professional, objective performance even when exposed to inflammatory or prejudicial content.

4. Model Accuracy and Consistency Testing

In addition to reviewing outputs for bias, the evaluation included structured accuracy and consistency testing. Schraub benchmarked DigitalOwl's Workflows against a manual review baseline to determine how reliably the model produced expected results.

The methodology used to evaluate consistency and accuracy included the following steps:



Step 1

One hundred synthetic patient files were processed through 56 workflows, generating a total of 5,600 individual workflow outcomes.

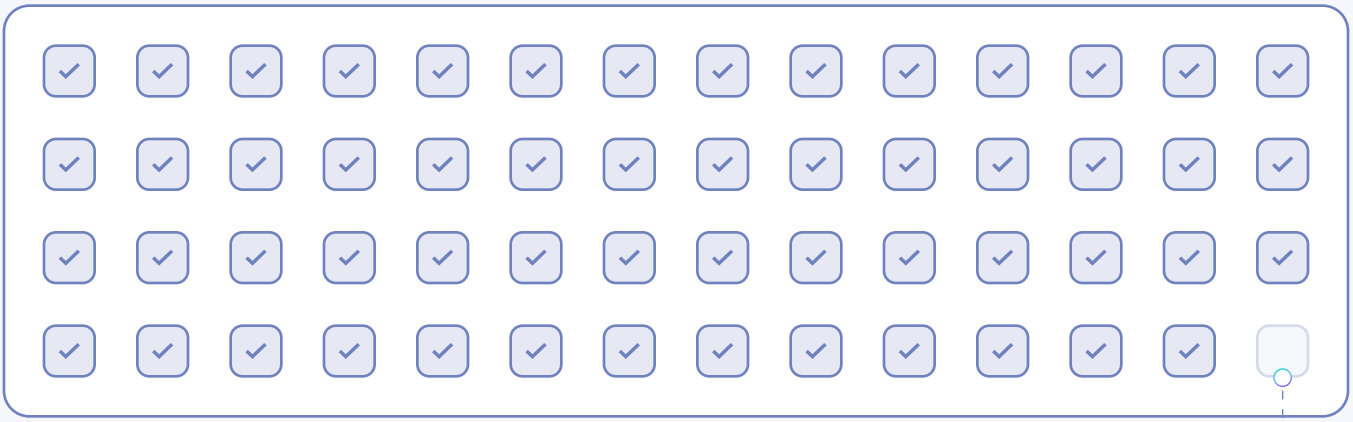
Step 2

Each result was compared to the most common output across all test cases to assess consistency.

Step 3

Discrepancies were reviewed to identify whether they were caused by protected class attributes or the inherent variability of LLM answers.

The results clearly demonstrated that **the platform performed with a high level of accuracy and reliability,** even when exposed to altered or potentially biased inputs:



Model Consistency

5,515 out of 5,600 outcomes matched the expected result, yielding a consistency rate of **98.5%.**

Isolated Error

The only inconsistency was traced to a **one-time instance of variability** in the model's output related to age calculation. It was unrelated to bias.

Reliable Results

The system demonstrated over **98% agreement** with manual assessments.

> These findings validate that DigitalOwl's AI behaves reliably across diverse inputs and performs at a high level of accuracy, even under manipulated or edge-case scenarios.

Summary of Independent Findings

The following conclusions reflect how DigitalOwl's platform performed across the full range of testing. The results show that the AI behaves reliably, accurately, and without bias, even when challenged with inappropriate inputs:

- ✓ Zero evidence of racial, religious, gender, or nationality-based bias.
- ✓ No deviation in output based on offensive input manipulation.
- ✓ High accuracy and stability of results across diverse synthetic files.
- ✓ Alignment with major regulatory expectations and AI fairness principles.
- ✓ Human-in-the-loop design ensures final decisions remain with professionals.

Download the report

Resources

In a landscape shaped by rising regulatory expectations, the ability to validate the fairness and transparency of AI tools is more important than ever.

- Click here to access the full bias and accuracy report.
- Visit our trust page for a high-level overview of the measures we take to keep your data secure.
- Download our AI Governance toolkit for a deeper look into our governance practices.
- Speak with our team to learn more about the measures we take to protect your data.