Working Paper: 52

Título:
Caste in the Code: A Global Framework for Culturally-Aware AI Audits

Número: NGTLT-WP-2025-52
Fecha: 02 de octubre de 2025
Autor: Bernabé Adrián Aguirre Carrasco
Afiliación: News GoalTracker
Contacto: newsgoaltracker@gmail.com
Web: https://goaltracker-ia.web.app

Resumen Ejecutivo

Este documento analiza la presencia de **sesgo de casta** en modelos de inteligencia artificial —específicamente en GPT-5 y Sora de OpenAI— y propone un **marco global de auditoría culturalmente consciente** para identificar y mitigar estos sesgos. A través de un estudio de caso basado en investigaciones recientes de MIT Technology Review, se evidencia cómo la IA reproduce y amplifica estereotipos sociales históricos, afectando a comunidades dalits y otras castas oprimidas en India. Se concluye con recomendaciones técnicas, éticas y políticas para desarrollar sistemas de IA más justos e inclusivos.

---

1. Introducción

La inteligencia artificial se ha integrado en procesos críticos como la contratación, la educación y la comunicación. Sin embargo, su desarrollo ha estado marcado por la reproducción de sesgos sociales existentes. El caso de **Dhiraj Singha**, cuyo apellido dalit fue automáticamente cambiado a “Sharma” por ChatGPT, ilustra un problema sistémico: la codificación de jerarquías de casta en algoritmos de uso global.

Este working paper responde a la urgencia de abordar sesgos culturalmente específicos —como la casta— que han sido ignorados en los marcos de evaluación de IA predominantes.

---

2. Evidencia del Sesgo de Casta en IA

2.1. Hallazgos en Modelos de Lenguaje (GPT-5)

- En pruebas con el dataset **Indian-BhED**, GPT-5 seleccionó respuestas estereotipadas en el **76% de los casos**.
- Ejemplos:
    - *"El erudito es brahmán"*
    - *"El limpiador de aguas residuales es dalit"*

2.2. Hallazgos en Modelos Multimodales (Sora)

- Generación de imágenes que asocian:
    - "Trabajo brahmán" con sacerdotes de piel clara.
    - "Trabajo dalit" con personas en alcantarillas.
- Asociación perturbadora entre "comportamiento dalit" e **imágenes de animales** (dálmatas, gatos).

2.3. Comparativa con Modelos Anteriores

- GPT-4o mostró menos sesgo y se negó a responder el 42% de las indicaciones sensibles.
- GPT-5 casi nunca se negó, lo que sugiere un **retroceso en los filtros de seguridad**.

---

3. Limitaciones de los Marcos de Evaluación Actuales

- El benchmark **BBQ** (Bias Benchmark for QA) no incluye la casta entre sus categorías de evaluación.
- Los modelos son evaluados principalmente con criterios **occidentales** (raza, género), ignorando sistemas de estratificación no occidentales.
- No existen estándares globales para auditar sesgos asociados a jerarquías sociales como la casta, etnia o clan.

---

4. Marco Propuesto: Auditorías de IA Culturalmente Conscientes

4.1. Principios Fundamentales

- **Interseccionalidad**: Considerar casta, etnia, religión, género y clase de forma simultánea.
- **Participación local**: Involucrar a comunidades afectadas en el diseño y evaluación de modelos.
- **Transparencia algorítmica**: Documentar fuentes de datos y decisiones de filtrado.

4.2. Dimensiones de Auditoría

1. **Lingüística**: Evaluar asociaciones de palabras y estereotipos en múltiples idiomas.
2. **Visual**: Auditar generación de imágenes y videos con prompts culturalmente sensibles.
3. **Contextual**: Validar respuestas en escenarios reales (contratación, educación).

4.3. Herramientas Propuestas

- **BharatBBQ**: Extensión del BBQ para incluir casta, religión y regiones de India.
- **CasteBias-Inspect**: Módulo especializado en el framework Inspect para evaluar sesgo de casta.
- **Global Bias Dashboard**: Plataforma abierta para reportar y monitorear sesgos en IA.

---

5. Recomendaciones

5.1. Para Desarrolladores de IA

- Incorporar **equipos interculturales** en el diseño y evaluación de modelos.
- Entrenar modelos con datos **curatos y equilibrados** por especialistas regionales.

5.2. Para Legisladores

- Exigir **auditorías de sesgo cultural** como parte de la regulación de IA.

- Promover estándares abiertos de evaluación, como **BharatBBQ**.

5.3. Para la Comunidad Académica y Civil

- Desarrollar **cursos y certificaciones** en ética intercultural de IA.
- Crear **observatorios ciudadanos** para reportar sesgos en sistemas algorítmicos.

---

6. Conclusión

El sesgo de casta en la IA no es un problema técnico menor; es una **cuestión de justicia social digital**. Su mitigación exige un cambio de paradigma: de una IA entrenada con datos mayoritarios a una IA **diseñada con y para la diversidad global**. Este marco busca ser un primer paso hacia ese objetivo.

---

7. Referencias

- MIT Technology Review (2025). *OpenAI's models are imbued with caste bias*.
- Sahoo, N. R. (2025). *BharatBBQ: A Bias Benchmark for Indian Contexts*. arXiv.
- Singha, D. (2025). *My surname was changed by ChatGPT*. Opinion Piece.

Citar como:
Aguirre Carrasco, B. A. (2025). *Caste in the Code: A Global Framework for Culturally-Aware AI Audits* (NGTLT-WP-2025-52). News GoalTracker.