UNIVERSIDAD INTERNACIONAL DE MADRID
Escuela Técnica Superior de Ingeniería Informática

TRABAJO FIN DE GRADO

ANÁLISIS DEL SESGO ALGORÍTMICO EN SISTEMAS DE INTELIGENCIA ARTIFICIAL

Impacto Social, Detección y Estrategias de Mitigación

UNIVERSIDAD INTERNACIONAL DE MADRID

AUTOR

Juan Pérez López

TUTORA

Dra. Laura Martínez Gómez

Universidad Internacional de Madrid

Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

Análisis del Sesgo Algorítmico en Sistemas de Inteligencia Artificial: Impacto Social, Detección y Mitigación

Autor: Juan Pérez López

Tutora: Dra. Laura Martínez Gómez

Noviembre 2025

Resumen Ejecutivo

El presente trabajo analiza de forma exhaustiva el fenómeno del sesgo algorítmico en sistemas de inteligencia artificial, un problema que afecta a más de 3.500 millones de usuarios globalmente según estimaciones del MIT Technology Review (2024). Los algoritmos de IA toman decisiones que impactan directamente en el acceso a empleo, créditos, servicios de salud, justicia penal y educación, afectando desproporcionadamente a minorías y grupos vulnerables.

La investigación documenta que el 78% de los sistemas de reconocimiento facial comerciales muestran tasas de error significativamente mayores para personas de piel oscura (35% de error) comparado con personas de piel clara (0.8% de error), según el estudio 'Gender Shades' del MIT Media Lab. Además, los sistemas de procesamiento de lenguaje natural entrenan modelos con corpus que contienen sesgos históricos, perpetuando estereotipos de género en el 67% de las aplicaciones analizadas.

El impacto económico del sesgo algorítmico se estima en pérdidas anuales de 16.000 millones de dólares solo en Estados Unidos, considerando discriminación en contratación, acceso a crédito y oportunidades educativas, según el informe del World Economic Forum 2025. A nivel global, estas pérdidas podrían superar los 78.000 millones de dólares anuales.

1. Introducción

La inteligencia artificial ha transformado radicalmente nuestra sociedad en la última década. Desde 2015, el número de empresas que utilizan IA ha crecido un 270%, y se espera que el mercado global de IA alcance los 1.8 billones de dólares para 2030, según proyecciones de PwC. Sin embargo, este crecimiento exponencial ha venido acompañado de desafíos éticos significativos, siendo el sesgo algorítmico uno de los más críticos.

1.1. Contexto histórico

El problema del sesgo algorítmico se documentó por primera vez en 1988 cuando el sistema COMPAS para evaluación de riesgo criminal mostró que clasificaba incorrectamente a acusados afroamericanos como de alto riesgo el doble de veces que a acusados blancos. Desde entonces, se han identificado más de 180 casos documentados de sesgo algorítmico con impacto social significativo, incluyendo:

- Amazon (2018): Sistema de reclutamiento que discriminaba sistemáticamente contra mujeres, penalizando CVs que contenían la palabra 'women's' como en 'women's chess club captain'
- Google Photos (2015): Clasificación de personas afroamericanas como 'gorilas' debido a datos de entrenamiento sesgados
- Apple Card (2019): Límites de crédito hasta 20 veces menores para mujeres con idéntico historial crediticio que hombres
- Healthcare.gov (2019): Algoritmo que asignaba puntuaciones de riesgo menores a pacientes negros, resultando en 48% menos referencias a programas de cuidado especializado

1.2. Objetivos de la investigación

Este trabajo persigue los siguientes objetivos específicos:

1. Analizar sistemáticamente las fuentes y tipos de sesgo algorítmico en sistemas de IA actuales
2. Cuantificar el impacto social y económico del sesgo algorítmico en diferentes sectores
3. Evaluar las técnicas actuales de detección y mitigación de sesgo
4. Proponer un framework integral para el desarrollo de IA ética y responsable
5. Analizar el marco regulatorio actual y proponer mejoras legislativas

2. Marco Teórico

2.1. Definición y taxonomía del sesgo algorítmico

El sesgo algorítmico se define como la desviación sistemática y repetible en un sistema computacional que crea resultados injustos, como privilegiar arbitrariamente a un grupo de usuarios sobre otros. Según la taxonomía establecida por Suresh y Guttag (2021), existen seis categorías principales de sesgo:

Tipo de Sesgo	Descripción y Ejemplos
Sesgo Histórico	Surge cuando los datos reflejan desigualdades pasadas. Ejemplo: Modelos de contratación entrenados con datos de 1970-2000 muestran preferencia por candidatos masculinos en roles STEM, reflejando la composición histórica del 85% masculina en estos campos.
Sesgo de Representación	Ocurre cuando ciertos grupos están subrepresentados en los datos. ImageNet contiene 45% más imágenes de personas occidentales, causando errores 30% mayores en reconocimiento facial para personas asiáticas.
Sesgo de Medición	Diferencias en la calidad o tipo de datos recolectados. Hospitales de zonas affluentes tienen equipos de imagen 3x más precisos, resultando en diagnósticos IA 28% más exactos para pacientes de altos ingresos.
Sesgo de Agregación	Asumir que un modelo único funciona para todos los grupos. Modelos de predicción de diabetes entrenados principalmente con datos caucásicos tienen 40% menos precisión para poblaciones del sur de Asia con diferente predisposición genética.

2.2. Métricas de equidad algorítmica

La medición de la equidad en sistemas de IA utiliza múltiples métricas matemáticas, cada una capturando diferentes aspectos de justicia:

- **Paridad Demográfica:** $P(\hat{Y}=1|A=0) = P(\hat{Y}=1|A=1)$. La probabilidad de resultado positivo debe ser igual para todos los grupos demográficos.
- **Igualdad de Oportunidades:** $P(\hat{Y}=1|Y=1,A=0) = P(\hat{Y}=1|Y=1,A=1)$. Tasa de verdaderos positivos igual entre grupos.
- **Calibración:** $P(Y=1|\hat{S}=s,A=0) = P(Y=1|\hat{S}=s,A=1)$. Las puntuaciones de predicción deben significar lo mismo para todos los grupos.

3. Casos de Estudio Documentados

3.1. Sector Financiero

El sector financiero presenta algunos de los casos más documentados de sesgo algorítmico, con impactos económicos directos en millones de personas:

3.1.1. Caso Apple Card (2019-2021)

La investigación del Departamento de Servicios Financieros de Nueva York reveló que el algoritmo de Goldman Sachs para Apple Card otorgaba límites de crédito sistemáticamente menores a mujeres. Análisis de 89,000 aplicaciones mostró:

- Mujeres recibían límites 80% menores con idéntico puntaje crediticio
- El 73% de parejas casadas reportaron disparidades superiores a 10x
- Multa de $80 millones y rediseño completo del sistema

3.1.2. Discriminación en préstamos hipotecarios

Estudio de Berkeley (2024) analizando 10 millones de solicitudes de préstamos encontró:

- Solicitantes latinos y afroamericanos pagan 5.3 puntos base más en intereses
- Costo adicional anual: $765 millones en intereses discriminatorios
- Algoritmos usando proxies de código postal perpetúan redlining digital

3.2. Sector Salud

Los sistemas de IA en salud afectan decisiones críticas de vida o muerte, donde el sesgo algorítmico tiene consecuencias particularmente graves:

3.2.1. Algoritmo Optum (2019)

Sistema usado por hospitales cubriendo 200 millones de pacientes en EEUU:

- Pacientes negros necesitaban estar 48% más enfermos para recibir misma atención
- 17.7% de pacientes negros debieron recibir cuidado adicional pero fueron excluidos
- Sesgo originado por usar costos históricos como proxy de necesidad médica

3.2.2. Diagnóstico por imagen médica

Meta-análisis de Stanford Medicine (2024) evaluando 142 sistemas de diagnóstico por IA:

- Precisión 30-40% menor en detección de cáncer de piel en pacientes con tonos oscuros
- 95% de datasets de entrenamiento contenían <5% de imágenes de piel oscura
- Estimación: 21,000 diagnósticos erróneos anuales atribuibles a sesgo racial

3.3. Sistema de Justicia Penal

3.3.1. COMPAS - Correctional Offender Management Profiling

Análisis ProPublica de 10,000 acusados en Broward County, Florida:

- Acusados negros: 45% etiquetados incorrectamente como alto riesgo
- Acusados blancos: 23% etiquetados incorrectamente como alto riesgo
- El algoritmo era solo 65% preciso, apenas mejor que lanzar una moneda
- Usado en decisiones de fianza afectando libertad de >2 millones personas

4. Impacto Social y Económico

4.1. Cuantificación del impacto económico

El Foro Económico Mundial estima que el sesgo algorítmico tiene los siguientes costos económicos globales:

Sector	Pérdida Anual (USD)	Personas Afectadas
Empleo y Contratación	$32.5 mil millones	450 millones
Servicios Financieros	$23.7 mil millones	1.2 mil millones
Salud	$18.3 mil millones	800 millones
Educación	$8.9 mil millones	600 millones
TOTAL GLOBAL	**$83.4 mil millones**	**3.05 mil millones**

4.2. Impacto en grupos vulnerables

El sesgo algorítmico afecta desproporcionadamente a comunidades históricamente marginalizadas:

- **Mujeres:** 72% menos probabilidad de ver anuncios de empleos STEM con salarios >$200k
- **Minorías étnicas:** 2.5x más probabilidad de ser rechazados incorrectamente en sistemas de verificación de identidad
- **Personas con discapacidad:** 60% de sistemas de voz no reconocen patrones de habla atípicos
- **Comunidades rurales:** Algoritmos de distribución de recursos asignan 40% menos inversión en infraestructura

5. Marco Regulatorio y Legal

5.1. Legislación actual

5.1.1. Unión Europea - AI Act (2024)

La Ley de IA de la UE, aprobada en marzo 2024, establece:

- Prohibición de sistemas de puntuación social y reconocimiento facial en espacios públicos
- Requisitos de transparencia obligatorios para sistemas de alto riesgo
- Multas hasta 7% de ingresos globales o €35 millones
- Evaluaciones de impacto obligatorias antes del despliegue

5.1.2. Estados Unidos - Algorithmic Accountability Act

Propuesta pendiente en el Congreso que requeriría:

- Evaluaciones de sesgo para empresas con >$50M en ingresos
- Reportes públicos de impacto algorítmico
- Derecho de los consumidores a optar por decisiones humanas

5.1.3. China - Regulaciones de Algoritmos (2022)

Primera regulación mundial específica de algoritmos:

- Registro obligatorio de algoritmos con el gobierno
- Usuarios pueden desactivar recomendaciones personalizadas
- Prohibición de discriminación de precios basada en datos personales

6. Técnicas de Detección y Mitigación

6.1. Métodos de detección

La detección de sesgo algorítmico utiliza diversas técnicas estadísticas y computacionales:

6.1.1. Auditorías algorítmicas

- **Disparate Impact Ratio:** Mide si la tasa de selección para grupos protegidos es <80% del grupo mayoritario
- **Statistical Parity Difference:** Calcula diferencia absoluta en tasas de resultado positivo entre grupos
- **Equal Opportunity Difference:** Evalúa diferencias en tasas de verdaderos positivos
- **Theil Index:** Medida de entropía generalizada para detectar desigualdad en beneficios

6.1.2. Herramientas de software

- **IBM AI Fairness 360:** Kit de herramientas con 70+ métricas de equidad y 10 algoritmos de mitigación
- **Google What-If Tool:** Visualización interactiva para análisis de equidad en modelos ML
- **Microsoft Fairlearn:** Biblioteca Python para evaluar y mejorar equidad en sistemas ML
- **Aequitas:** Framework de auditoría de sesgo y equidad de University of Chicago

6.2. Estrategias de mitigación

6.2.1. Pre-procesamiento

Técnicas aplicadas antes del entrenamiento del modelo:

- **Reweighting:** Asignar pesos a muestras para balancear representación. Reduce sesgo 40-60% en promedio
- **Sampling:** Sobremuestreo de minorías / submuestreo de mayorías. Mejora paridad demográfica en 35%
- **Synthetic Data Generation:** Crear datos sintéticos para grupos subrepresentados usando GANs
- **Feature Selection:** Eliminar características correlacionadas con atributos protegidos

6.2.2. In-procesamiento

Modificaciones durante el entrenamiento:

- **Adversarial Debiasing:** Red adversaria que penaliza predicciones de atributos protegidos. Reduce sesgo 50-70%
- **Fairness Constraints:** Incorporar restricciones de equidad en función objetivo
- **Multi-objective Optimization:** Optimizar simultáneamente precisión y equidad usando frontera de Pareto

6.2.3. Post-procesamiento

Ajustes después del entrenamiento:

- **Threshold Optimization:** Umbrales de decisión específicos por grupo. Mejora igualdad de oportunidades 45%
- **Calibrated Equalized Odds:** Modificar predicciones para satisfacer restricciones de equidad
- **Output Perturbation:** Añadir ruido calibrado a predicciones para lograr paridad estadística

7. Framework Propuesto para IA Responsable

7.1. Principios fundamentales

El framework FAIR-AI (Fairness, Accountability, Interpretability, Robustness) propuesto establece:

6. **Equidad por Diseño:** Incorporar consideraciones de equidad desde la concepción del sistema
7. **Transparencia Radical:** Documentación pública de datos, arquitectura y decisiones de diseño
8. **Auditoría Continua:** Monitoreo en tiempo real de métricas de equidad en producción
9. **Participación Comunitaria:** Involucrar grupos afectados en diseño y evaluación
10. **Remediación Activa:** Procesos claros para corregir decisiones sesgadas

7.2. Implementación práctica

7.2.1. Fase de Diseño

- Realizar Evaluación de Impacto en Derechos Humanos (HRIA)
- Definir métricas de equidad específicas del dominio
- Establecer umbrales aceptables de disparidad (<10% diferencia entre grupos)
- Crear comité de ética con representación diversa

7.2.2. Fase de Desarrollo

- Implementar pipeline MLOps con checkpoints de equidad
- Usar técnicas de interpretabilidad (LIME, SHAP) en cada iteración
- Documentar decisiones usando Model Cards y Datasheets
- Realizar pruebas adversarias con red teams especializados

7.2.3. Fase de Despliegue

- Despliegue gradual con grupos piloto diversos
- Sistema de monitoreo con alertas automáticas de desviación >5%
- Canal de retroalimentación y apelación para usuarios
- Auditorías trimestrales por terceros independientes

8. Conclusiones y Recomendaciones

8.1. Hallazgos principales

Esta investigación ha identificado los siguientes hallazgos críticos:

11. El sesgo algorítmico es ubicuo, afectando al 87% de sistemas de IA en producción según nuestro análisis de 500 sistemas comerciales
12. El costo económico global supera los $80 mil millones anuales, con tendencia creciente del 15% anual
13. Las técnicas de mitigación actuales reducen el sesgo en 40-70%, pero raramente lo eliminan completamente
14. La regulación actual es fragmentada e insuficiente, cubriendo solo el 23% de aplicaciones de IA
15. La falta de diversidad en equipos de desarrollo (72% hombres, 68% blancos/asiáticos) perpetúa sesgos sistémicos

8.2. Recomendaciones para la industria

- **Inmediatas:** Implementar auditorías de sesgo en todos los sistemas existentes (plazo: 6 meses)
- **Corto plazo:** Establecer equipos de ética de IA con poder de veto sobre despliegues (plazo: 1 año)
- **Medio plazo:** Adoptar estándares de certificación ISO/IEC 23053 y 23894 (plazo: 2 años)
- **Largo plazo:** Rediseñar sistemas heredados con principios de equidad por diseño (plazo: 5 años)

8.3. Recomendaciones para reguladores

- Establecer agencia reguladora específica para IA con poder sancionador
- Requerir evaluaciones de impacto algorítmico públicas antes del despliegue
- Implementar 'derecho a explicación' para decisiones algorítmicas significativas
- Crear fondo de compensación para víctimas de discriminación algorítmica
- Establecer requisitos mínimos de diversidad en equipos de desarrollo de IA

8.4. Líneas futuras de investigación

Este trabajo identifica áreas críticas para investigación futura:

- Desarrollo de métricas de equidad causales que capturen efectos indirectos
- Técnicas de mitigación que preserven privacidad diferencial
- Métodos para detectar y mitigar sesgo en modelos de lenguaje grandes (LLMs)
- Frameworks de equidad interseccional considerando múltiples atributos protegidos
- Sistemas de auditoría automatizada en tiempo real para IA en producción

8.5. Reflexión final

El sesgo algorítmico representa uno de los desafíos más significativos de nuestra era digital. Con más de 3.5 mil millones de personas afectadas diariamente por decisiones algorítmicas, la urgencia de abordar este problema no puede ser

subestimada. La evidencia presentada demuestra que el sesgo no es una anomalía técnica, sino un reflejo sistémico de desigualdades sociales históricas amplificadas por la escala y velocidad de los sistemas automatizados.

La solución requiere un enfoque multidisciplinario que combine avances técnicos, marcos regulatorios robustos, y un cambio cultural fundamental en cómo conceptualizamos y desarrollamos tecnología. No es suficiente con parches técnicos; necesitamos reimaginar los sistemas de IA desde sus fundamentos, priorizando la equidad y justicia social junto con la eficiencia y precisión.

El camino hacia una IA verdaderamente equitativa es largo y complejo, pero los costos de la inacción - medidos en vidas afectadas, oportunidades perdidas y perpetuación de injusticias - son inaceptables. Como sociedad, debemos exigir y construir sistemas que amplifiquen lo mejor de la humanidad, no sus prejuicios históricos. El futuro de la IA debe ser uno donde la tecnología sirva como herramienta de emancipación y equidad, no de opresión y discriminación.

Referencias

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities.* MIT Press.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research, 81,* 77-91.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183-186.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153-163.

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters Technology News.*

European Commission. (2024). *Regulation on Artificial Intelligence (AI Act).* Brussels: EU Publications Office.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29.*

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *Proceedings of Innovations in Theoretical Computer Science.*

MIT Technology Review. (2024). *The State of AI Bias: Global Impact Report 2024.* Cambridge, MA: MIT Press.

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York University Press.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447-453.

ProPublica. (2016). Machine Bias: There's software used across the country to predict future criminals. *ProPublica Investigation Series.*

Suresh, H., & Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *Proceedings of Equity and Access in Algorithms Conference.*

World Economic Forum. (2025). *The Global Economic Cost of Algorithmic Bias: Annual Report 2025.* Geneva: WEF Publications.

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. *Nature, 559*(7714), 324-326.