ÉTICA Y GÉNERO EN LA IA: IDENTIFICAR SESGOS DE GÉNERO EN IA MEDIANTE PENSAMIENTO COMPLEJO

Fecha de recepción: 16-05-24 / Fecha de aceptación: 16-06-24

María Nely Vásquez Pérez
Universidad de Deusto

Correo: marianely.vasquez@deusto.es
ORCID: https://orcid.org/0000-0002-0879-5309

Resumen

El presente trabajo explora la identificación y mitigación de sesgos aplicando categorías propias del Pensamiento Complejo, especialmente los sesgos de género, en los modelos de inteligencia artificial (IA), así como las mejores prácticas para garantizar la equidad y la inclusión en el desarrollo de algoritmos de IA.

Palabras Clave: Inteligencia Artificial, Identificación de Sesgos Algorítmicos, Sesgos de Género, Pensamiento Complejo.

Abstract:

This work explores the identification and mitigation of biases by applying categories from Complex Thinking, especially gender biases, in artificial intelligence (AI) models, as well as best practices to ensure fairness and inclusion in the development of AI algorithms.

Keywords: Artificial Intelligence, Identification of Algorithmic Biases, Gender Biases, Complex Thinking.

1. Introducción.

En el mundo de la tecnología de vanguardia, la inteligencia artificial (IA) se ha convertido en un motor de innovación y transformación. Esta tecnología ha revolucionado diversas industrias, desde la medicina hasta el transporte, mejorando la eficiencia y creando nuevas oportunidades. Sin embargo, a medida que su adopción se generaliza, también lo hace el reconocimiento de sus limitaciones y desafíos, particularmente en términos de sesgo algorítmico y equidad. El sesgo en los modelos de IA puede perpetuar e incluso exacerbar las desigualdades sociales, lo que hace imperativo desde el punto de vista de la Ética de la IA abordar estos problemas de manera proactiva. Especialmente evidente y grave (por su generalización) es la aparición de sesgos algorítmicos discriminatorios por razón de género.

El sesgo algorítmico puede surgir de diversas fuentes, como datos de entrenamiento no representativos o prejuicios inherentes en los algoritmos. Estos sesgos pueden tener consecuencias significativas, como la discriminación en la contratación laboral, la justicia penal y el acceso a servicios financieros, o la mencionada por razón de género. Por ello, es crucial desarrollar estrategias para identificar y mitigar estos sesgos. Incorporar categorías de análisis del Pensamiento Complejo y las Ciencias de la Complejidad puede suponer una contribución decisiva para la identificación de dichos sesgos. Estas disciplinas permiten una comprensión más holística y multifacética de los sistemas de IA, considerando las interacciones y retroalimentaciones dentro de estos sistemas.

Además, la colaboración interdisciplinaria entre expertos en IA, científicos sociales y éticos puede proporcionar enfoques más integrales y efectivos para abordar los desafíos éticos. La transparencia en el diseño de algoritmos y la participación de diversas comunidades en el proceso de desarrollo de IA son pasos importantes hacia la creación de sistemas más justos y equitativos. En resumen, para que la IA sea verdaderamente transformadora y beneficiosa para toda la sociedad, es fundamental abordar sus limitaciones y trabajar hacia la equidad y la justicia algorítmica.

2. Identificación y mitigación de Sesgos en Modelos de IA

2.1. Identificación

a) Fuentes de Sesgo

Datos Sesgados: Uno de los principales orígenes del sesgo en IA proviene de los conjuntos de datos utilizados para entrenar los modelos. Si estos datos no representan de manera justa a todos los grupos de personas, el modelo resultante probablemente perpetuará o amplificará estos sesgos (Mehrabi et al. 2021). Este problema se agrava cuando los datos reflejan prejuicios históricos o sociales, lo que puede llevar a decisiones injustas en áreas como la contratación laboral, la justicia penal y el acceso a servicios financieros.

Diseño del Algoritmo: Además, los algoritmos pueden diseñarse, intencionada o accidentalmente, de manera que favorezcan ciertos resultados. La transparencia y la comprensión del proceso de toma de decisiones son cruciales para identificar y mitigar sesgos de diseño (Holstein et al. 2019). Es esencial que los desarrolladores de IA comprendan cómo sus decisiones en el diseño del algoritmo pueden influir en los resultados y trabajen activamente para evitar estas injusticias.

b) Métodos de Detección

Auditorías de Equidad: Las auditorías de equidad se refieren a un proceso de evaluación meticuloso y sistemático de los algoritmos, con el propósito de detectar posibles desviaciones que podrían indicar la presencia de sesgos. Este tipo de auditorías se ha vuelto fundamental en el contexto de la inteligencia artificial y el aprendizaje automático, dado que estos sistemas se utilizan cada vez más en decisiones críticas que afectan a personas y comunidades. El objetivo principal es asegurar que los algoritmos sean justos y equitativos, minimizando así el riesgo de discriminación basada en características como raza, género, edad, o cualquier otra variable sensible.

Un estudio destacado en este campo es el de Raji y Buolamwini, quienes han señalado la importancia de estas auditorías para identificar y mitigar sesgos en los sistemas de reconocimiento facial. Su investigación demostró que muchos de estos sistemas tienden a funcionar de manera desigual para diferentes grupos demográficos, lo que subraya la necesidad de implementar auditorías de equidad como una práctica estándar en el desarrollo y despliegue de tecnologías basadas en algoritmos (Raji y Buolamwini 2019).

Análisis de Sensibilidad: El análisis de sensibilidad, por otro lado, es una técnica utilizada para evaluar cómo las variaciones en los datos de entrada pueden influir en los resultados de un modelo. Este método es crucial para identificar dependencias indebidas en ciertas características, lo que puede revelar puntos débiles en el diseño del algoritmo y su capacidad de generalización.

Es un método que ayuda a los desarrolladores y científicos de datos a entender mejor el comportamiento de sus modelos bajo diferentes escenarios y condiciones. Al modificar sistemáticamente los valores de entrada y observar los cambios en los resultados, se puede determinar si un modelo está sobreajustado a ciertas características específicas o si es capaz de manejar una diversidad de datos de manera robusta.

Por ejemplo, en aplicaciones de crédito, un análisis de sensibilidad puede revelar si el modelo depende excesivamente de variables como el código postal, lo cual podría indirectamente introducir sesgos geográficos. Este tipo de análisis es esencial para asegurar que los modelos sean justos y que las decisiones automatizadas basadas en ellos no perpetúen desigualdades existentes (Binns, 2018).

2.2. Mitigación de Sesgos en IA

a) Limpieza de Datos

La limpieza de datos es una fase crucial en el proceso de desarrollo de modelos de inteligencia artificial y aprendizaje automático. Este paso inicial se centra en asegurar que los datos de entrenamiento sean lo más inclusivos y representativos posible, lo cual es esencial para mitigar el sesgo en los modelos. La calidad de los datos de entrada determina en gran medida la equidad y la eficacia de los algoritmos resultantes (Wang et al. 2019).

Para lograr una limpieza de datos efectiva, se deben considerar varias estrategias. Una de las más importantes es la recolección de datos adicionales de grupos subrepresentados. En muchos casos, los conjuntos de datos originales pueden estar sesgados debido a una falta de diversidad en las muestras recolectadas. Por ejemplo, un conjunto de datos de imágenes podría tener una sobre-representación de personas de cierto grupo étnico o género, lo que lleva a que el modelo entrenado sobre estos datos tenga un rendimiento desigual cuando se aplica a un grupo más diverso.

Además, la limpieza de datos no solo se refiere a la inclusión de diversos grupos demográficos, sino también a la eliminación de datos ruidosos o irrelevantes que podrían introducir sesgos inadvertidos. Esto incluye la detección y corrección de errores en los datos, la eliminación de duplicados y la normalización de las variables para asegurar consistencia.

Por ejemplo, en el contexto de un sistema de contratación automatizada, es fundamental que los datos de entrenamiento incluyan una variedad de perfiles de candidatos de diferentes géneros, edades, etnias y antecedentes educativos. Si los datos de entrenamiento solo incluyen candidatos de un determinado grupo demográfico, el modelo resultante probablemente perpetuará ese sesgo, favoreciendo a los mismos tipos de candidatos en el futuro.

b) Diseño de Algoritmos Consciente

El diseño de algoritmos consciente es un enfoque que busca minimizar los sesgos y mejorar la equidad en los sistemas de inteligencia artificial y aprendizaje automático. Este enfoque incluye varias estrategias, entre las cuales destacan la transparencia y la regularización.

Transparencia: La transparencia en el diseño de algoritmos es fundamental para asegurar que tanto el funcionamiento del algoritmo como sus decisiones sean comprensibles para los humanos (Dwork et al. 2012). Ello implica que los procesos internos del algoritmo deben ser accesibles y claros para los desarrolladores, usuarios y cualquier parte interesada. Esto es crucial para identificar y corregir posibles fuentes de sesgo y para generar confianza en la tecnología.

Para lograr transparencia, se pueden implementar varias prácticas. Una de ellas es la documentación detallada del proceso de desarrollo del algoritmo, incluyendo las decisiones tomadas durante el diseño y el entrenamiento. Esta documentación debe explicar cómo

se seleccionaron los datos de entrenamiento, qué técnicas se utilizaron para preprocesar estos datos y cómo se ajustaron los parámetros del modelo.

Otra práctica importante es el uso de algoritmos interpretables, como los árboles de decisión, que permiten a los usuarios entender cómo se llega a una determinada decisión.

Regularización: Las técnicas de regularización son esenciales para evitar que un modelo dependa demasiado de características específicas que podrían ser fuentes de sesgo. La regularización ayuda a mejorar la generalización del modelo, reduciendo el riesgo de sobreajuste y asegurando que las predicciones sean más equitativas y robustas.

Una técnica que se aplica es la inclusión de términos de equidad en la función de pérdida del modelo. Esto implica ajustar la función de pérdida para penalizar las predicciones que muestran signos de sesgo, incentivando al modelo a producir resultados más equilibrados.

Por ejemplo, en un sistema de recomendación de empleo, la regularización podría ayudar a evitar que el modelo dependa demasiado de características como la universidad de procedencia o el historial laboral, que podrían favorecer injustamente a ciertos candidatos sobre otros. Al aplicar técnicas de regularización adecuadas, el sistema puede aprender a valorar una gama más amplia de habilidades y experiencias, promoviendo una selección de candidatos más justa y diversa.

c) Pruebas Continuas y Monitoreo

Las pruebas continuas y el monitoreo son componentes esenciales en el ciclo de vida de los modelos de inteligencia artificial y aprendizaje automático. Estas prácticas garantizan que los modelos no solo sean precisos y eficaces en su implementación inicial, sino que también mantengan su equidad y rendimiento a lo largo del tiempo.

Evaluación Continua

La evaluación continua de los modelos es crucial para asegurar su rendimiento a largo plazo (Mitchell et al. 2019). La evaluación no debe ser un evento único que ocurre solo durante la fase de desarrollo. En su lugar, debe ser un proceso continuo que incluya pruebas regulares de equidad y precisión. Este enfoque permite detectar y corregir problemas que puedan surgir debido a cambios en los datos de entrada, variaciones en las condiciones de operación o el surgimiento de nuevos patrones de uso.

Para implementar una evaluación continua, se pueden establecer sistemas automáticos de monitoreo que realicen pruebas periódicas en el modelo utilizando conjuntos de datos de validación actualizados. Estas pruebas deben incluir métricas específicas de equidad para asegurar que el modelo no esté introduciendo o amplificando sesgos. Por ejemplo, en un modelo de recomendación de productos, las pruebas podrían incluir verificaciones de que las recomendaciones no favorezcan sistemáticamente a ciertos grupos de usuarios sobre otros.

Retroalimentación de los Usuarios

La incorporación de la retroalimentación de los usuarios finales es otra estrategia clave para identificar y corregir sesgos que pueden no haber sido evidentes durante el desarrollo del modelo. Los usuarios pueden proporcionar información valiosa sobre cómo el modelo se comporta en situaciones del mundo real y señalar casos en los que las predicciones o decisiones del modelo parecen injustas o sesgadas.

Para facilitar la retroalimentación de los usuarios, se pueden implementar interfaces y canales de comunicación que permitan a los usuarios reportar problemas fácilmente. Esta retroalimentación debe ser recopilada y analizada de manera sistemática para identificar patrones y áreas de mejora. Por ejemplo, en una aplicación de préstamo automatizada, los usuarios pueden reportar casos en los que sienten que sus solicitudes fueron injustamente rechazadas. Estos informes pueden ser analizados para identificar y corregir posibles fuentes de sesgo en el modelo.

2.3. Mejores Prácticas para la Equidad y la Inclusión

a) Desarrollo Inclusivo

El desarrollo inclusivo es un enfoque esencial en la creación de modelos de inteligencia artificial y aprendizaje automático que busca minimizar los sesgos y promover la equidad desde las primeras etapas del diseño. Este enfoque incluye la formación de equipos de desarrollo diversos y la participación activa de las partes interesadas.

Equipos Diversos

La diversidad en los equipos de desarrollo es fundamental para garantizar que los modelos de IA sean justos y equitativos. Equipos de desarrollo diversificados pueden aportar una amplia gama de perspectivas y experiencias, lo que es crucial para identificar y mitigar sesgos desde el inicio del proceso de desarrollo. La inclusión de personas con diferentes antecedentes culturales, étnicos, de género y profesionales puede enriquecer la discusión y el análisis crítico de las decisiones de diseño.

Los equipos diversos son más capaces de anticipar y abordar los posibles impactos negativos de los modelos de IA en diferentes comunidades. Por ejemplo, en el desarrollo de un sistema de reconocimiento facial, la diversidad del equipo puede ayudar a identificar problemas de precisión que afectan desproporcionadamente a ciertos grupos demográficos, como las personas de piel más oscura. Al reconocer y abordar estos problemas desde el principio, los desarrolladores pueden crear modelos más robustos y equitativos.

Participación de las Partes Interesadas

Incluir a las partes interesadas, especialmente a aquellas de grupos subrepresentados, en el proceso de desarrollo es otra estrategia clave para asegurar la equidad en los modelos de IA. La participación activa de estas partes interesadas (como los consumidores o usuarios) puede proporcionar valiosas perspectivas y conocimientos que los desarrolladores pueden no tener.

La participación de las partes interesadas puede tomar diversas formas, como consultas, talleres colaborativos y pruebas piloto. Estas interacciones permiten a los desarrolladores obtener retroalimentación directa sobre cómo los modelos de IA afectan a diferentes grupos y ajustar sus diseños en consecuencia. Por ejemplo, en el desarrollo de una aplicación de salud digital, la participación de pacientes de diversas comunidades puede revelar preocupaciones específicas sobre privacidad, accesibilidad y relevancia cultural que deben ser abordadas para asegurar la equidad y eficacia del sistema.

Además, la inclusión de partes interesadas en el proceso de desarrollo fomenta la transparencia y la responsabilidad. Al involucrar a representantes de los grupos afectados, los desarrolladores pueden demostrar su compromiso con la equidad y ganar la confianza de las comunidades a las que sirven. Esto también puede ayudar a identificar y mitigar posibles riesgos éticos y sociales asociados con la implementación de tecnologías de IA (West et al. 2019).

b) Marcos Regulatorios y Éticos

Los marcos regulatorios y éticos son fundamentales para asegurar que el desarrollo y la implementación de la inteligencia artificial (IA) se realicen de manera justa, equitativa y responsable. Estos marcos proporcionan directrices y normas que ayudan a mitigar riesgos, proteger los derechos de los individuos y promover la equidad en el uso de tecnologías avanzadas.

Principios Éticos

Adoptar principios éticos claros que guíen el desarrollo de la IA es esencial para asegurar un enfoque centrado en la equidad (Jobin et al. 2019).

Los principios éticos sirven como una brújula moral que orienta a los desarrolladores y organizaciones en la toma de decisiones durante el diseño, desarrollo y despliegue de sistemas de IA. Los principios éticos más básicos deben incluir:

1. **Justicia y Equidad:** Asegurar que los sistemas de IA no perpetúen ni amplifiquen las desigualdades existentes y que sus beneficios se distribuyan de manera equitativa entre todas las comunidades.
2. **Transparencia y Explicabilidad:** Garantizar que los procesos y decisiones de los sistemas de IA sean comprensibles y accesibles para los usuarios y las partes interesadas.

3. **Responsabilidad:** Establecer mecanismos claros para la rendición de cuentas, de modo que los desarrolladores y operadores de IA puedan ser responsabilizados por las consecuencias de sus sistemas.
4. **Privacidad y Seguridad:** Proteger la privacidad de los datos personales y asegurar que los sistemas de IA sean robustos frente a ataques y manipulaciones.
5. **Beneficencia:** Asegurar que los sistemas de IA se desarrollen y utilicen con el objetivo de mejorar el bienestar humano y social.

Implementar estos principios éticos requiere un compromiso continuo y una reflexión crítica sobre las implicaciones de la IA en diversos contextos.

Cumplimiento Regulatorio

El cumplimiento regulatorio es igualmente importante para asegurar que los sistemas de IA operen dentro de los límites legales y protejan los derechos de los individuos. Estar al tanto de y cumplir con las regulaciones locales e internacionales relacionadas con la IA y la no discriminación es crucial para evitar infracciones legales y promover la confianza en las tecnologías de IA.

Las regulaciones pueden variar ampliamente entre diferentes jurisdicciones, pero suelen incluir aspectos como:

1. **Protección de Datos:** Leyes como el Reglamento General de Protección de Datos (GDPR) en la Unión Europea imponen estrictos requisitos sobre cómo se deben recopilar, almacenar y procesar los datos personales.
2. **No Discriminación:** Legislaciones que prohíben la discriminación basada en raza, género, edad, orientación sexual, discapacidad y otras características protegidas. Estas leyes requieren que los sistemas de IA no introduzcan ni perpetúen sesgos discriminatorios.
3. **Transparencia Algorítmica:** Algunas regulaciones exigen que las organizaciones revelen cómo funcionan sus algoritmos y permitan la supervisión y auditoría externa para asegurar la equidad y la transparencia.
4. **Seguridad y Resiliencia:** Normas que establecen requisitos para garantizar la seguridad y resiliencia de los sistemas de IA contra fallos y ataques cibernéticos.

Cumplir con estas regulaciones no solo ayuda a evitar sanciones legales, sino que también fortalece la confianza del público en las tecnologías de IA. Las organizaciones deben man-

tener una vigilancia constante sobre los desarrollos regulatorios y adaptar sus prácticas conforme a las nuevas leyes y directrices.

3. Identificación y mitigación de Sesgos de Género en Modelos de IA

3.1. Identificación de Sesgos Algorítmicos de Género

a) Fuentes de Sesgo

Datos Sesgados: Un ejemplo clásico es el uso de datos históricos para entrenar modelos de IA en la contratación o promoción laboral (Bolukbasi et al. 2016). Si estos datos reflejan una historia de predominio masculino en ciertas posiciones, el modelo puede aprender incorrectamente que los hombres son preferibles para estos roles, perpetuando así la desigualdad de género.

Diseño del Algoritmo: Consideremos los asistentes de voz, muchos de los cuales han sido programados con voces femeninas para roles serviciales (Hupfer et al. 2020). Este diseño puede reforzar estereotipos de género sobre los roles de las mujeres en la sociedad.

b) Métodos de Detección

Auditorías de Equidad: Al analizar cómo los modelos de IA toman decisiones que afectan a diferentes géneros, podemos identificar sesgos (Buolamwini y Gebru 2018). Por ejemplo, un análisis de las recomendaciones de empleo por parte de un sistema de IA podría revelar una preferencia injustificada por candidatos de un género sobre otro.

Análisis de Sensibilidad: Cambiar el género de los perfiles de usuarios en los datos de prueba y observar cómo varían las recomendaciones de un sistema (De-Arteaga et al. 2019) puede exponer dependencias indebidas en el género.

c) Mitigación de Sesgos de Género en IA

Limpieza de Datos

Se debe hacer un esfuerzo consciente para asegurar que los datos de entrenamiento sean equilibrados en términos de género. Esto puede implicar aumentar la representación de géneros subrepresentados en los datos o aplicar técnicas de ponderación para equilibrar la influencia de los ejemplos de entrenamiento (Zhao et al. 2017).

Diseño de Algoritmos Consciente

Regularización de Equidad: Una técnica fundamental es implementar técnicas de regularización que penalizan al modelo por tomar decisiones basadas en el género de manera indebida (Hardt et al. 2016).

Modelos Explicables: Desarrollar modelos que puedan explicar sus decisiones con transparencia, por ejemplo mediante diagramas de árboles o esquemas de ramificación de decisiones permite entender el funcionamiento del modelo e identificar y corregir posibles sesgos de género.

Pruebas Continuas y Monitoreo

Se deben establecer procedimientos para evaluar regularmente los modelos de IA en busca de sesgos de género, incluso después de su implementación (Raji y Yang 2019), para asegurar que las mejoras en la equidad sean sostenibles a largo plazo.

d) Ejemplos Concretos de Aplicación

Hay numerosos ejemplos de grandes empresas que han corregido sus algoritmos tras la identificación de sesgos de género en los mismos. Amazon desechó una herramienta de selección de personal basada en IA después de descubrir que favorecía a los candidatos masculinos para roles técnicos. Este caso subraya la importancia de la vigilancia continua y el análisis de equidad en los sistemas de IA (Dastin, 2018).

También tras el monitoreo de diversas herramientas de traducción se han evidenciado sesgos al traducir de idiomas neutrales en cuanto al género a idiomas con género gramatical, asignando roles profesionales a hombres y roles de soporte a mujeres (Saunders y Byrne 2020). La mitigación aquí implica ajustar o corregir los algoritmos para ofrecer alternativas neutrales o equilibradas en cuanto al género.

Combatir el sesgo de género en la IA requiere un enfoque multifacético que aborde tanto la calidad de los datos como el diseño y la implementación de algoritmos. A través de la identificación proactiva de sesgos, el diseño inclusivo y ético, y un compromiso continuo con la evaluación y mejora, podemos avanzar hacia sistemas de IA que promuevan la equidad de género y contrarresten las desigualdades existentes. La clave está en reconocer que la tecnología no es neutral y que las decisiones que tomamos en su desarrollo y despliegue pueden tener un impacto significativo en la promoción de una sociedad más justa e inclusiva.

4. Identificación de Sesgos de Género aplicando Pensamiento Complejo al análisis de los algoritmos

Suele reconocerse sin mucho problema que la Inteligencia Artificial no es una mera herramienta, así como tampoco es una mera ciencia y, mucho menos, una mera rama de la computación, sino más bien un conjunto de ciencias muy diversas que funcionan de ma-

nera interdisciplinaria. Este es el dato fundamental para definir la IA como Ciencia de la Complejidad. ¿Podrá entonces el Pensamiento Complejo aportar categorías de análisis que facilite la identificación de sesgos algorítmicos, contribuyendo a un desarrollo ético y equitativo de la IA? Naturalmente que sí.

4.1. Reconocimiento de la Complejidad de los Datos

Los datos son fundamentales para el funcionamiento de los sistemas de IA, pero también son una fuente primordial de sesgo. El pensamiento complejo nos lleva a examinar críticamente los conjuntos de datos no solo por su contenido explícito sino también por lo que omiten. En el caso del sesgo de género, esto significa preguntar quiénes están representados en los datos, quiénes no lo están y cómo se presentan estas representaciones (Buolamwini y Gebru 2018). Por ejemplo, si los conjuntos de datos para el reconocimiento facial están desproporcionadamente compuestos por rostros masculinos blancos, los sistemas entrenados con estos datos tendrán dificultades para reconocer rostros femeninos y de personas de color, reflejando y perpetuando sesgos existentes.

4.2. Análisis de las Interacciones entre Algoritmos y Sociedad

El pensamiento complejo también aboga por analizar cómo los sistemas de IA interactúan con las estructuras sociales existentes. Los sesgos de género en la IA no solo surgen de los datos y algoritmos sino también de cómo estos sistemas se utilizan y los efectos que tienen en la sociedad (Noble, 2018). Por ejemplo, si un algoritmo de contratación prioriza características que históricamente han sido asociadas con candidatos masculinos, esto no solo refleja sesgos en los datos de entrenamiento sino que también puede reforzar las desigualdades de género en el lugar de trabajo. De hecho, una de las categorías de análisis crítico poscolonial, la interseccionalidad, es prácticamente equivalente a complejidad en cuanto a las interacciones o intersecciones entre diversos sistemas y subsistemas (Fricker 2007).

4.3. Consideración de la Diversidad y la Inclusión en el Desarrollo de la IA

Desde la perspectiva del pensamiento complejo, la diversidad y la inclusión en los equipos de desarrollo de IA son fundamentales para identificar y mitigar sesgos. Equipos diversos pueden aportar una gama más amplia de experiencias y perspectivas, lo que es esencial para reconocer suposiciones implícitas y sesgos potenciales en el diseño y la implementación de sistemas de IA (West et al. 2019). Este enfoque también subraya la importancia de incluir voces marginadas y subrepresentadas en el proceso de desarrollo, lo que puede ayudar a anticipar y abordar problemas de sesgo de género antes de que los sistemas se desplieguen.

4.4. Estrategias para la Mitigación de Sesgos de Género

a) Desarrollo de Marcos Éticos y Regulatorios

El pensamiento complejo aboga por la creación de marcos éticos y regulatorios que guíen el desarrollo y la implementación de la IA. Estos marcos deben reconocer la complejidad de los sesgos de género y proporcionar principios y directrices claros para abordarlos. Esto incluye la implementación de normas para la recopilación y el uso de datos, así como la evaluación continua de los sistemas de IA en busca de sesgos y discriminación (Jobin et al. 2019).

b) Fomento de la Transparencia y la Responsabilidad

Para combatir eficazmente los sesgos de género, los sistemas de IA deben ser transparentes en cuanto a cómo funcionan y cómo toman decisiones. El pensamiento complejo enfatiza la necesidad de mecanismos que permitan a los usuarios y a las partes interesadas comprender y cuestionar las decisiones de IA (Doshi-Velez y Kim 2017). Esto está estrechamente ligado a la responsabilidad, donde los desarrolladores y las empresas que despliegan sistemas de IA deben ser responsables de los impactos de sus tecnologías.

c) Promoción de la Educación y la Concienciación

Finalmente, el pensamiento complejo destaca la importancia de la educación y la concienciación sobre los sesgos de género en la IA. Esto incluye la capacitación de los profesionales de IA en temas de ética y sesgo, así como la sensibilización del público sobre cómo los sistemas de IA pueden perpetuar desigualdades (Eubanks 2018). La educación y la concienciación son pasos cruciales para fomentar un diálogo más amplio sobre la equidad y la inclusión en la tecnología.

5. Conclusión

La Inteligencia Artificial (IA) es una ciencia interdisciplinaria que requiere un enfoque de pensamiento complejo para identificar sesgos algorítmicos, especialmente los sesgos de género, y fomentar su desarrollo ético y equitativo. Ello es debido a que el pensamiento complejo nos lleva a examinar críticamente los datos no solo por su contenido explícito, sino también por lo que omiten, preguntando quiénes están representados y quiénes no.

Además, este enfoque analiza cómo los sistemas de IA interactúan con las estructuras sociales, identificando cómo los sesgos de género no solo surgen de los datos y algoritmos, sino también de su uso y efectos en la sociedad.

La diversidad e inclusión en unos equipos de desarrollo de IA que sean profundamente interdisciplinares (otra de las características del pensamiento complejo) son esenciales para reconocer y mitigar sesgos. Equipos diversos aportan una gama más amplia de experiencias y perspectivas, lo que ayuda a anticipar y abordar problemas de sesgo de género antes de que surjan.

Referencias

- *Binns, R. (2018). "Fairness in machine learning: Lessons from political philosophy". Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency, 149-159.*
- *Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings". Advances in neural information processing systems, 29.*
- *Buolamwini, J., & Gebru, T. (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification". Proceedings of the 1st Conference on Fairness, Accountability, and Transparency, 77-91.*
- *Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters.*
- *De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., ... & Geyik, S. (2019). "Bias in bios: A case study of semantic representation bias in a high-stakes setting". Proceedings of the Conference on Fairness, Accountability, and Transparency, 120-128.*
- *Doshi-Velez, F., & Kim, B. (2017). "Towards a rigorous science of interpretable machine learning". arXiv:1702.08608v2.*
- *Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). "Fairness through awareness". Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214-226.*
- *Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press.*
- *Fricker, M. (2007). Epistemic Injustice. Power and the Ethics of Knowing. Oxford Univesity Press.*
- *Hardt, M., Price, E., & Srebro, N. (2016). "Equality of opportunity in supervised learning". Advances in neural information processing systems, 29, 3315-3323.*
- *Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). "Improving fairness in machine learning systems: What do industry practitioners need?". Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-16.*
- *Hupfer, S., O'Rourke, E., Park, J. S., Young, M., & Choi, J. (2020). "The Gendered Design of AI Assistants: Speaking, Serving, and Gender Stereotypes". Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1-13.*
- *Jobin, A., Ienca, M., & Vayena, E. (2019). "The global landscape of AI ethics guidelines". Nature Machine Intelligence, 1(9), 389-399.*
- *Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). "A survey on bias and fairness in machine learning". ACM Computing Surveys (CSUR), 54(6), 1-35.*
- *Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). "Model cards for model reporting". Proceedings of the Conference on Fairness, Accountability, and Transparency, 220-229.*
- *Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. NYU Press.*
- *Raji, I. D., & Buolamwini, J. (2019). "Action-*

Referencias

able auditing: Investigating the impact of publicly naming biased performance results of commercial AI products". Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 429-435.

- *Raji, I. D., & Yang, X. (2019). "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing". Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33-44.*
- *Saunders, L., & Byrne, B. (2020). "Reducing gender bias in neural machine translation as a domain adaptation problema". Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 772-782.*
- *Wang, Y., Wu, L., & Wang, H. (2019). "Mitigating bias in facial recognition datasets". 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2761-2768.*
- *West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race, and power in AI. AI Now Institute.*
- *Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). "Men also like shopping: Reducing gender bias amplification using corpus-level constraints". Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2979-2989.*