

A collage of office workers in various settings, including a man looking at a laptop, a woman smiling, a man with headphones, and a woman working on a laptop.

PRODUCTO 3

**GUÍA DE BUENAS PRÁCTICAS PARA
IDENTIFICACIÓN, PREVENCIÓN Y MITIGACIÓN
DE SESGOS DE GÉNERO EN EL DESARROLLO Y
ADOPCIÓN DE LOS SISTEMAS DE IA**

CONTRATO DNP-1104-2024

Bogotá D.C. Diciembre de 2024





Departamento Nacional de Planeación (DNP)
Directora de Desarrollo Digital (DDD)
Viviana Rocío Vanegas Barrero

Coordinadora Grupo de Transformación y Economía Digital (DDD)
Yenifer Julie Pinto Gaitán

Comité Técnico de Seguimiento
Sara Daniela Márquez Gutiérrez
Martha Magdalena Sánchez Rodríguez
Hugo Marlon Arenas Domínguez
Juan Sebastián Numpaque Cano

Beta Group Colombia
Equipo Consultor
Anderson Danilo Betancourt Betancourt – Director proyecto
Armando Guio Español – Experto en ética IA
Diego Fernando Cristancho - Investigador
Karen Lizeth Amezcua Rodríguez - Investigadora
Maria Antonia Carvajal – Experta en ética IA
Marianela Luzardo Briceño – Experta en IA
Olga Cecilia Ramírez Roa – Experta en género

Cítese este documento así:
DNP. (2024). GUÍA DE BUENAS PRÁCTICAS PARA LA IDENTIFICACIÓN,
PREVENCIÓN Y MITIGACIÓN DE SESGOS DE GÉNERO EN EL DESARROLLO Y
ADOPCIÓN DE LOS SISTEMAS DE IA. Departamento Nacional de Planeación
(DNP). Bogotá: Beta Group Colombia

Nota de transparencia:

Este documento ha sido elaborado con la asistencia de ChatGPT, un modelo de lenguaje basado en inteligencia artificial desarrollado por OpenAI, específicamente utilizando la arquitectura GPT-4. El modelo ha sido empleado para mejorar la claridad, coherencia y estructura del texto. Sin embargo, todas las ideas, contenido y decisiones finales son responsabilidad exclusiva de los autores. La utilización de esta herramienta no ha afectado la integridad ni el rigor de la información presentada.



Tabla de Contenido

5

1. Introducción

8

2. Principales Sesgos en Sistemas de IA

10

3. Buenas Prácticas para el Desarrollo de los Sistemas de IA

18

4. Buenas Prácticas para la Adopción de los Sistemas de IA

20

5. Catálogo de Herramientas y Árbol de Decisión

21 5.1 Catálogo de herramientas

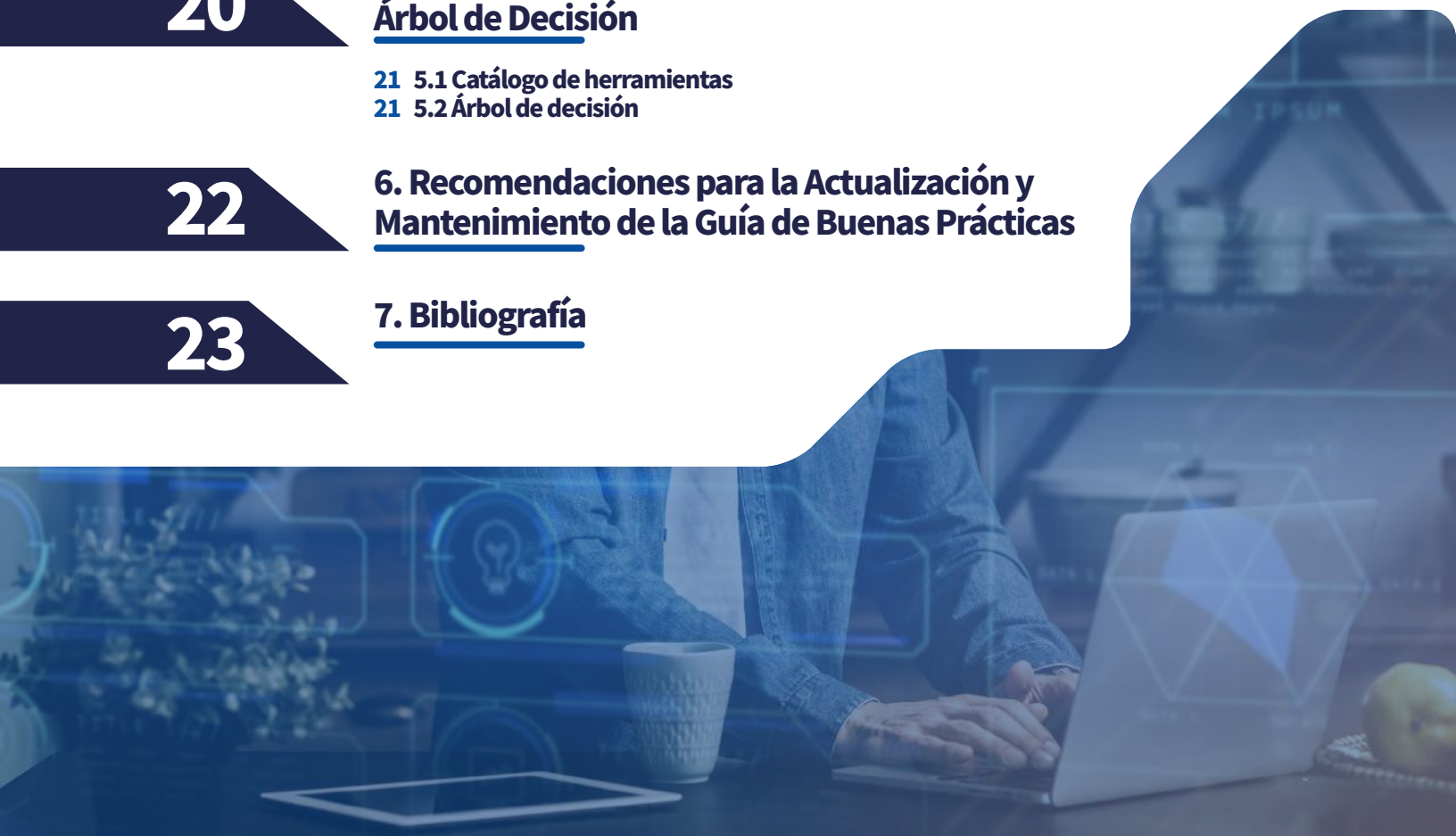
21 5.2 Árbol de decisión

22

6. Recomendaciones para la Actualización y Mantenimiento de la Guía de Buenas Prácticas

23

7. Bibliografía



Introducción

1

La inteligencia artificial (IA) es una tecnología transformadora que está redefiniendo la manera en que personas, organizaciones y gobiernos abordan desafíos en múltiples sectores. Sin embargo, su diseño e implementación deben realizarse de manera ética para evitar que perpetúe o amplifique desigualdades preexistentes, particularmente en términos de género. Por ello, resulta crucial establecer estrategias claras que permitan identificar, prevenir y mitigar los sesgos de género en el desarrollo y uso de sistemas de IA.

En este contexto, el CONPES 4080 que establece la Política Pública de Equidad de Género para las Mujeres en Colombia, destaca la equidad de género como un pilar esencial del desarrollo sostenible. Dentro de este marco, el Departamento Nacional de Planeación (DNP) lideró la elaboración de la presente guía de buenas prácticas para mitigar los sesgos de género en la IA, promoviendo principios éticos y de justicia social (CONPES, 2022).

Esta guía fue desarrollada por BETA GROUP COLOMBIA S.A.S. para el DNP, bajo el contrato DNP-1104-2024 adjudicado mediante concurso de méritos. Su propósito es facilitar herramientas para fomentar la equidad de género en Colombia, abarcando tanto el ecosistema como el ciclo de vida de los sistemas de IA, con un enfoque en la inclusión y la responsabilidad ética.

Con un enfoque integral, la guía aborda todas las etapas del ciclo de vida de la IA, desde la planificación y el diseño hasta el monitoreo y el desmantelamiento, promoviendo sistemas inclusivos. También incorpora un Catálogo de Herramientas y un Árbol de Decisión como recursos prácticos para identificar y mitigar sesgos de género, facilitando su aplicación en diversos contextos.

Para facilitar su implementación, la guía está dirigida a una diversidad de personas o entidades clave que desempeñan roles esenciales en el ecosistema de la IA, entre los que se encuentran:



Equipos de desarrollo de IA, quienes tienen la responsabilidad de diseñar tecnologías inclusivas desde la fase inicial.



Grupos de defensa de género, que pueden contribuir a garantizar que los datos sean representativos y éticos.



Académicos e investigadores, encargados de avanzar en el conocimiento y la innovación técnica con un enfoque inclusivo.



Empresas tecnológicas, responsables de implementar sistemas que respeten los principios de equidad y ética.



Reguladores y organismos gubernamentales, que supervisan el uso responsable y alineado con las políticas públicas.



Personas usuarias finales, quienes deben ser protegidas de los impactos negativos y participar activamente en los procesos de mejora continua.



Medios de comunicación y periodistas tecnológicos, como actores clave en la sensibilización y comunicación de los impactos sociales de la IA.

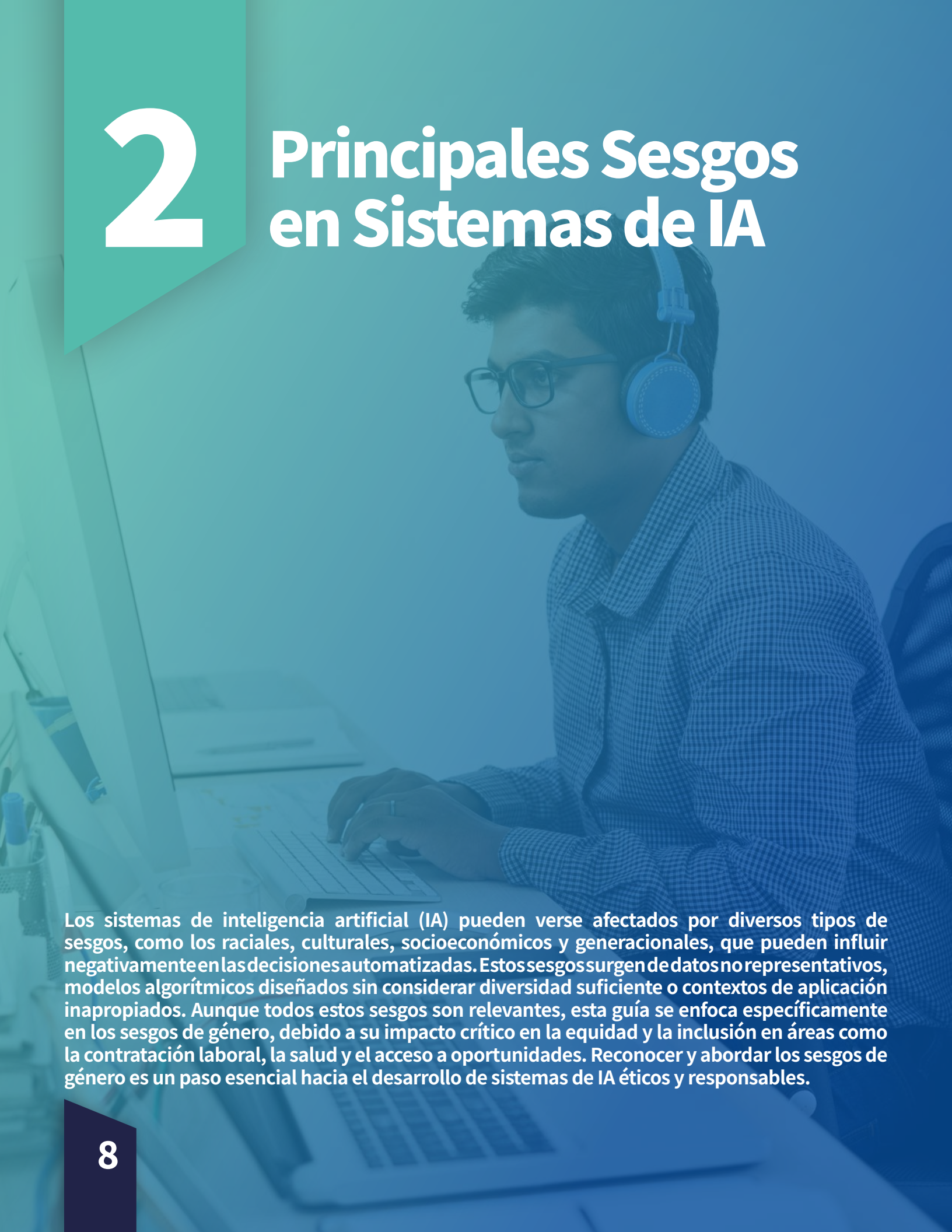
Introducción

El cumplimiento de esta acción reafirma el compromiso de Colombia con la adopción de marcos éticos y recomendaciones internacionales para el uso responsable de la inteligencia artificial. Con ejemplos concretos y soluciones estructuradas, esta guía busca convertir principios éticos en acciones prácticas, fomentando el desarrollo de sistemas de IA técnicamente avanzados y socialmente responsables. Este enfoque asegura que la IA se utilice como una herramienta para promover la equidad, la inclusión y el respeto a los principios fundamentales de ética y justicia social, en línea con los estándares establecidos por organismos internacionales como la OCDE y la UNESCO.



2

Principales Sesgos en Sistemas de IA

A man with dark hair, wearing glasses and large blue headphones, is seated at a desk. He is looking at a computer monitor and has his hands on a keyboard. The background is a soft-focus office environment. The entire image has a blue and green color overlay.

Los sistemas de inteligencia artificial (IA) pueden verse afectados por diversos tipos de sesgos, como los raciales, culturales, socioeconómicos y generacionales, que pueden influir negativamente en las decisiones automatizadas. Estos sesgos surgen de datos no representativos, modelos algorítmicos diseñados sin considerar diversidad suficiente o contextos de aplicación inapropiados. Aunque todos estos sesgos son relevantes, esta guía se enfoca específicamente en los sesgos de género, debido a su impacto crítico en la equidad y la inclusión en áreas como la contratación laboral, la salud y el acceso a oportunidades. Reconocer y abordar los sesgos de género es un paso esencial hacia el desarrollo de sistemas de IA éticos y responsables.

Tabla 1. Principales sesgos en sistemas de IA

| Tipo de Sesgos | Descripción |
|-------------------------------------|---|
| Sesgo de datos | Se presenta cuando los algoritmos se entrenan con datos no representativos, excluyendo las experiencias de mujeres y personas LGBTQ+, lo cual genera decisiones sesgadas. En el ámbito de la salud, esto afecta diagnósticos y tratamientos al no considerar diferencias fisiológicas entre géneros diversos, (Buolamwini & Gebru, 2018); en el ámbito laboral, refuerza la preferencia por perfiles masculinos, perpetuando la exclusión de mujeres y minorías en ciertos roles históricos (O'Neil, 2016). |
| Sesgo de confirmación | Este sesgo ocurre cuando los algoritmos priorizan información que refuerza estereotipos preexistentes, como asociar roles laborales tradicionalmente masculinos, lo cual perpetúa estas ideas preconcebidas en procesos de contratación y selección de personal (Buolamwini & Gebru, 2018). |
| Sesgo de interacción | La interacción de los sistemas de IA varía según el género, reflejando y reforzando estereotipos tradicionales en las respuestas de asistentes virtuales. Por ejemplo, los sistemas de IA como Alexa y Siri utilizan voces femeninas en roles serviles y voces masculinas en contextos autoritarios, lo cual contribuye a la idea tradicional de género asociada a la autoridad y el servicio (UNESCO, 2024). |
| Sesgo de uso | Las barreras estructurales dificultan el acceso equitativo de mujeres y personas LGBTQ+ a los sistemas de IA. La falta de acceso a dispositivos tecnológicos y la exclusión en los datos de entrenamiento limitan las oportunidades de estos grupos para beneficiarse plenamente de los avances tecnológicos, lo que amplía las brechas de género existentes (UNESCO, 2024). |
| Sesgo de retroalimentación | Se produce cuando los resultados de un sistema de IA influyen en los datos futuros que procesará el mismo sistema, creando un ciclo que refuerza los sesgos preexistentes. Este ciclo amplifica las desigualdades en áreas como la contratación, donde se perpetúa la exclusión de las mujeres, o en salud, al no tener en cuenta sus necesidades específicas en los tratamientos y diagnósticos (Raji & Buolamwini, 2019). |
| Sesgo de contextual o de aplicación | Aparece cuando los sistemas de IA se aplican en contextos diferentes a los previstos, generando discriminación de género al no ajustarse a las necesidades socioculturales específicas de mujeres y personas LGBTQ+. Los algoritmos que no consideran diversas realidades socioculturales terminan exacerbando desigualdades y limitando la efectividad de los sistemas en diferentes entornos (O'Neil, 2016). |

Fuente: Consultoría BetaGroup

Los distintos tipos de sesgos de género descritos evidencian cómo los sistemas de inteligencia artificial pueden perpetuar desigualdades al reflejar limitaciones inherentes en sus datos, diseños o aplicaciones. El abordaje de estas problemáticas es esencial para garantizar que la implementación de la AI sea inclusiva y representativa, mitigando el impacto negativo que estos sesgos pueden tener en diversos ámbitos. Con este contexto, se han categorizado las buenas prácticas en función de los tipos de sesgos que pueden surgir en cada etapa del ciclo de vida de la IA.

3

Buenas Prácticas para el Desarrollo de los Sistemas de IA

Los sistemas de inteligencia artificial (IA) tienen el potencial de transformar sectores clave y contribuir al progreso social y económico. Sin embargo, es fundamental garantizar que su desarrollo y uso promuevan la equidad de género, evitando la reproducción de desigualdades existentes y fortaleciendo su impacto positivo en individuos y comunidades.

Para abordar estos desafíos, es importante establecer buenas prácticas que guíen la identificación, prevención y mitigación de sesgos de género en todas las etapas del ciclo de vida de los sistemas de IA, desde su planificación y diseño hasta su desmantelamiento. Las siguientes prácticas se fundamentan en recomendaciones éticas de organismos internacionales como la UNESCO (2021) y la OCDE (2024) quienes subrayan la importancia de fomentar la equidad, diversidad e inclusión en el desarrollo de tecnologías de IA.

Buenas Prácticas para el Desarrollo de los Sistemas de IA

La implementación de buenas prácticas en la planificación y diseño de la IA es esencial para mitigar sesgos de género desde el inicio, garantizar inclusión y equidad, y reflejar valores éticos y sociales. Una planificación sólida y directrices claras previenen desigualdades estructurales, fomentando tecnologías justas y beneficiosas para todos.

Tabla 2. Buenas prácticas en la etapa de planificación y diseño

| Descripción / Enfoque | Ejemplo de Aplicación |
|---|---|
| <p>Implementar directrices inclusivas en el diseño para asegurar la representación justa y precisa de las comunidades de género diverso, incluidas las personas LGBTQ+, en todas las fases del desarrollo de productos y servicios impulsados por IA.</p> <p>Enfoque: Prevención de sesgos de género.</p> | <p>GLAAD, en colaboración con Google, ha creado “All In”, un conjunto de recursos de marketing inclusivo que ofrece directrices detalladas para representar de manera auténtica a comunidades subrepresentadas, incluidas las personas LGBTQ+. Estas directrices buscan ayudar a los profesionales del marketing a evitar estereotipos y a reflejar con precisión la diversidad de la comunidad LGBTQ+ (GLAAD, 2021).</p> |
| <p>Establecer metas de equidad de género claras y medibles en el diseño de IA, asegurando que los resultados de los sistemas de IA respeten los principios de no discriminación y diversidad.</p> <p>Enfoque: Identificación y prevención de sesgos de género.</p> | <p>Accenture ha establecido metas ambiciosas para lograr una fuerza laboral con equilibrio de género para el 2025. Esto significa una fuerza laboral que sea mitad mujeres y mitad hombres para personas con género binario (Accenture, s.f.).</p> |
| <p>Definir previamente las prácticas de diseño que permitan la transparencia en las decisiones algorítmicas y faciliten la comprensión de cómo los datos y modelos afectan las recomendaciones y resultados del sistema.</p> <p>Enfoque: Identificación de sesgos de género.</p> | <p>Google desarrolló el What-If Tool dentro de su suite de IA para mejorar la transparencia y explicabilidad en los modelos de aprendizaje automático. Esta herramienta permite a los desarrolladores analizar cómo variables de entrada, como género o edad, afectan las salidas del modelo, identificando si existen disparidades o sesgos en los resultados (Google, s.f.).</p> <p>Mediante la visualización de cómo las variables impactan las predicciones, los equipos pueden realizar ajustes en los modelos para tomar decisiones más justas. Esta práctica permite identificar y mitigar posibles sesgos de género desde la fase de diseño, promoviendo sistemas de IA más equitativos y confiables.</p> |
| <p>Incorporar una perspectiva de prevención de violencia de género digital desde las primeras etapas del diseño de sistemas de IA, asegurando que los productos no faciliten ni amplifiquen dinámicas de poder desiguales.</p> <p>Enfoque: Identificación y mitigación de riesgos de violencia de género</p> | <p>El documento de la UNFPA sobre la violencia de género facilitada por la tecnología (TF GBV) presenta directrices claras que abordan cómo la tecnología puede influir en la seguridad y bienestar de las mujeres y personas de identidades diversas, además de destacar las acciones necesarias para mitigar este tipo de violencia. Los principios clave incluyen la necesidad de diseño seguro y privacidad por defecto, que están estrechamente relacionados con la protección de los derechos de las personas y la seguridad digital (UNFPA, s.f.).</p> |

Fuente: Consultoría BetaGroup

La recopilación y tratamiento de datos en IA son cruciales para evitar la amplificación de sesgos y promover la equidad. Es clave garantizar datos diversos, representativos y validados, proteger la privacidad, e implementar auditorías y técnicas innovadoras que identifiquen y mitiguen sesgos desde el inicio. Estas prácticas aseguran sistemas de IA éticos, confiables y alineados con principios de equidad.

Tabla 3. Buenas prácticas en la etapa de recopilación y tratamiento de datos

| Descripción / Enfoque | Ejemplo de Aplicación |
|--|--|
| <p>Asegurar que los conjuntos de datos utilizados reflejen una diversidad de género adecuada y que no contengan sesgos que puedan reproducir estereotipos. Esto puede lograrse mediante auditorías de equidad algorítmica de los datos para evaluar su representatividad y calidad.</p> <p>Enfoque: Prevención de sesgos de género.</p> | <p>Empresas tecnológicas globales como Google, Microsoft e IBM han implementado políticas para reducir los sesgos de género en sus sistemas de IA mediante auditorías algorítmicas y directrices internas que buscan evitar desigualdades. Google, por ejemplo, realiza auditorías de equidad algorítmica, analizando datos desagregados por género, raza y otros factores para identificar y corregir posibles tendencias discriminatorias (Google AI, 2020).</p> |
| <p>Documentar los orígenes y las características de los conjuntos de datos, incluyendo cómo se abordan las cuestiones de género en la recopilación y procesamiento de datos.</p> <p>Enfoque: Prevención de sesgos de género.</p> | <p>Un caso de uso denominado “CelebA Dataset” fue publicado por la Universidad China de Hong Kong (CUHK) muestra un ejemplo de prevención de sesgos a través de la documentación detallada de datasets. Este conjunto de datos de celebridades incluye más de 200,000 imágenes con anotaciones detalladas, donde se documenta explícitamente la distribución de género, los métodos de recopilación y etiquetado, y los atributos relacionados con el género. La documentación incluye metadatos específicos sobre la composición demográfica, el proceso de recolección de imágenes, y cómo se abordaron las cuestiones de género en el etiquetado, haciendo que este dataset sea ampliamente utilizado en investigaciones sobre equidad y sesgos en sistemas de reconocimiento facial (Ziwei, Ping, Xiaogang, & Xiaoou, 2015).</p> |
| <p>Fomentar el uso de conjuntos de datos abiertos y estándares que garanticen diversidad de género, y establecer controles para minimizar la recolección de datos sesgados.</p> <p>Enfoque: Prevención de sesgos de género.</p> | <p>IBM desarrolló el conjunto de datos Diversity in Faces para fomentar la diversidad y reducir el sesgo en sistemas de reconocimiento facial. Este conjunto de datos, diseñado con una amplia representación de género, etnicidad y otros atributos demográficos, permite a los desarrolladores entrenar modelos de IA más inclusivos y representativos. Al promover estándares de diversidad y establecer controles para evitar la recolección de datos sesgados, IBM busca asegurar que los modelos de IA reconozcan correctamente a personas de diversos grupos (Exponiendo.ai, 2019).</p> |
| <p>Proteger la privacidad y anonimizar los datos de forma que se minimicen riesgos para las personas.</p> <p>Enfoque: Prevención de sesgos de género.</p> | <p>Apple emplea privacidad diferencial para proteger la información personal en aplicaciones como Siri y su sistema de autocorrección. Esta técnica agrega “ruido” a los datos recolectados, asegurando que la información individual se mantenga anónima incluso cuando se analizan patrones de uso general. Además, Apple limita el procesamiento de datos al dispositivo del usuario, lo cual significa que la mayoría de los datos no se envían a sus servidores. De este modo, Apple puede mejorar sus servicios y algoritmos de manera segura, minimizando los riesgos para la privacidad de las personas y asegurando que su información se mantenga protegida (Apple, s.f.).</p> |

Fuente: Consultoría BetaGroup

Buenas Prácticas para el Desarrollo de los Sistemas de IA

La creación y/o adaptación de modelos de IA son etapas críticas para prevenir la integración de sesgos de género en los algoritmos. Aplicar prácticas como técnicas de entrenamiento equitativas, pruebas exhaustivas y una cultura de responsabilidad en los equipos de desarrollo garantiza modelos justos y respetuosos con la diversidad. Estas estrategias permiten abordar los sesgos desde el núcleo, integrando principios éticos y herramientas innovadoras que los detectan y mitigan antes del despliegue, promoviendo una IA inclusiva y confiable.

Tabla 4. Buenas prácticas en la etapa de creación de modelo(s) y/o adaptación de modelo(s)

| Descripción / Enfoque | Ejemplo de Aplicación |
|--|--|
| Aplicar técnicas de entrenamiento y ajuste de modelos que consideren la equidad de género, utilizando algoritmos que limiten el sesgo algorítmico y se adapten para minimizar los efectos de género en las predicciones. Enfoque: Prevención de sesgos de género. | Google aplicó técnicas de entrenamiento y ajuste en Google Translate para reducir el sesgo de género en sus traducciones automáticas. Antes, en idiomas sin pronombres de género específicos, el sistema tendía a traducir con estereotipos, como “doctor” en masculino y “enfermera” en femenino. Para solucionar esto, Google ajustó el modelo de IA para que proporcione traducciones en ambos géneros cuando el contexto es ambiguo. Este enfoque ayuda a limitar el sesgo algorítmico y asegura que las traducciones reflejen una perspectiva más equitativa, ofreciendo opciones de traducción tanto en masculino como en femenino (Google, 2018). |
| Incluir etapas de validación y prueba enfocadas en detectar sesgos de género, utilizando métricas específicas que evalúen la justicia y la equidad de los resultados según el género y permitan ajustes antes del despliegue. Enfoque: Mitigación de sesgos de género. | Microsoft ha creado un equipo interno dedicado a la ética de la IA, con el objetivo de garantizar que sus productos cumplan con los principios de justicia y equidad. El enfoque de Microsoft incluye la creación de herramientas que permiten a los desarrolladores identificar y corregir sesgos en las fases tempranas del desarrollo de sistemas de IA. Una de estas herramientas es Fairlearn, que evalúa los modelos de IA para asegurar que las predicciones sean justas para todos los grupos sociales, incluidas las mujeres (Raji & Buolamwini, 2019). |
| Instruir a los equipos de desarrollo y responsables de IA en la identificación y mitigación de sesgos de género en los modelos, promoviendo una cultura de responsabilidad en la creación de IA. Enfoque: Identificación, prevención y mitigación de sesgos de género. | Google implementa el programa Machine Learning Fairness para capacitar a sus desarrolladores en la identificación y mitigación de sesgos en IA, incluyendo los de género. En este programa, los equipos aprenden a ajustar modelos en productos como Google Photos y Google Translate para evitar resultados estereotipados o discriminatorios. Por ejemplo, en Google Photos, los desarrolladores aplican técnicas para ajustar el reconocimiento de imágenes y evitar la categorización basada en estereotipos de género, y en Google Translate se ofrecen traducciones en ambos géneros en contextos ambiguos para reducir el sesgo (Google, s.f.) |
| Realizar una evaluación temprana para identificar riesgos de sesgos de género potenciales en el diseño de los sistemas de IA. Enfoque: Identificación de sesgos de género. | International Business Machines Corporation (IBM) ha desarrollado herramientas como AI Fairness 360 y Finspector para evaluar y mitigar los sesgos de género en los sistemas de inteligencia artificial. Estas herramientas permiten a los desarrolladores identificar y corregir sesgos en los modelos de IA, promoviendo la equidad y la justicia en sus aplicaciones (IBM, s.f.). |

Fuente: Consultoría BetaGroup

La etapa de prueba, evaluación, verificación y validación es clave para garantizar que los sistemas de IA sean justos, transparentes y libres de sesgos de género. Las buenas prácticas incluyen identificar y corregir desigualdades, validar tanto el desempeño técnico como la equidad de los modelos, y promover la inclusión de expertos interdisciplinarios y comunidades afectadas. Estas estrategias aseguran sistemas éticos, responsables y respetuosos con la diversidad antes de su despliegue.

Tabla 5. Buenas prácticas en la etapa de prueba, evaluación, verificación y validación

| Descripción / Enfoque | Ejemplo de Aplicación |
|---|--|
| <p>Realizar auditorías éticas y algorítmicas en las que participen profesionales de diversas disciplinas para analizar los impactos de género en los sistemas de IA, verificando que cumplan con los principios y objetivos de equidad.</p> <p>Enfoque: Identificación, prevención y mitigación de sesgos de género.</p> | <p>AI Now Institute, una organización de investigación líder en el análisis de las implicaciones sociales de la IA, ha publicado numerosos informes sobre la necesidad de auditar los sistemas de IA para mitigar los sesgos de género. Sus investigaciones han influido en la adopción de buenas prácticas por parte de las empresas tecnológicas, y su enfoque interdisciplinario incluye recomendaciones para asegurar que las auditorías algorítmicas sean obligatorias en las organizaciones que desarrollan IA (Whittaker, y otros, 2018).</p> |
| <p>Garantizar que los resultados de IA sean comprensibles y que los usuarios puedan cuestionar decisiones o resultados, especialmente aquellos que afectan a grupos de género subrepresentados.</p> <p>Enfoque: Mitigación de sesgos de género.</p> | <p>Google aplicó técnicas de explicabilidad en Google Translate para reducir el sesgo de género en sus traducciones automáticas, después de identificar que el sistema tendía a traducir ciertas palabras de manera estereotipada (por ejemplo, “doctor” como masculino y “enfermera” como femenino). Para abordar este problema, Google ajustó su modelo de traducción para que, cuando exista ambigüedad de género, se ofrezcan opciones tanto en masculino como en femenino. Esto no solo permite que los usuarios elijan la traducción adecuada, sino que también fomenta un uso inclusivo del sistema y ayuda a evitar sesgos de género (Google, 2018).</p> |
| <p>Realizar pruebas que simulen escenarios diversos para asegurar que el sistema responda de manera equitativa a una variedad de contextos de género.</p> <p>Enfoque: Mitigación de sesgos de género.</p> | <p>Una empresa destacada en este campo es Clever AI, un startup que ha diseñado modelos de IA con un enfoque inclusivo. Clever AI utiliza datos desagregados por género y realiza pruebas en diferentes grupos demográficos para asegurarse de que sus algoritmos funcionen de manera equitativa en diversos contextos (AI4ALL, 2020).</p> |
| <p>Involucrar a expertos en estudios de género, defensores de la equidad y a las comunidades afectadas en el desarrollo y evaluación de la IA.</p> <p>Enfoque: Identificación, prevención y mitigación de sesgos de género.</p> | <p>Algorithmic Justice League, ha desempeñado un papel clave en destacar los sesgos raciales y de género en los sistemas de IA. Esta organización trabaja con empresas tecnológicas para ayudarles a identificar y corregir sesgos en sus productos de IA. La liga también ha sido fundamental en la creación de conciencia pública sobre los riesgos que presentan los algoritmos sesgados y ha impulsado políticas para que las empresas tecnológicas adopten una mayor transparencia en sus procesos de desarrollo. (League, s.f.)</p> |
| <p>Garantizar la transparencia sobre cómo funciona la IA y hacer que los procesos de toma de decisiones sean explicables, para poder identificar y corregir sesgos de género.</p> <p>Enfoque: Identificación y mitigación de sesgos de género.</p> | <p>IBM Watson OpenScale permite realizar evaluaciones de equidad en modelos de IA, facilitando la identificación y corrección de sesgos de género y otros sesgos. Esta herramienta permite a los desarrolladores monitorear atributos como “Sexo” o “Edad” y configurar umbrales de equidad para comparar los resultados entre diferentes grupos, garantizando que los modelos no favorezcan injustamente a un género sobre otro. Además, Watson OpenScale aplica ajustes en tiempo real mediante métodos de atenuación de sesgo, asegurando que los modelos operen de manera justa y transparente mientras están en uso (IBM, 2024).</p> |

Fuente: Consultoría BetaGroup

Buenas Prácticas para el Desarrollo de los Sistemas de IA

La etapa de despliegue es crítica para garantizar que los sistemas de IA mantengan equidad y transparencia al interactuar con los usuarios. Las buenas prácticas incluyen monitoreo continuo, canales de retroalimentación y ajustes dinámicos según el uso real, asegurando sistemas inclusivos que identifiquen y mitiguen sesgos emergentes. Estas estrategias refuerzan la confianza y promueven la justicia en su aplicación.

Tabla 6. Buenas prácticas en la etapa de entrada en servicio/despliegue

| Descripción / Enfoque | Ejemplo de Aplicación |
|---|--|
| Implementar un sistema de monitoreo continuo que permita detectar sesgos de género en la operación de IA, especialmente a medida que el sistema interactúa con distintos grupo de usuarios. Enfoque: Identificación y mitigación de sesgos de género. | IBM ha sido pionera en la creación de una plataforma de IA ética, que incluye herramientas para mitigar sesgos y promover la transparencia. IBM ha lanzado la plataforma AI Fairness 360, un conjunto de bibliotecas y algoritmos que ayudan a los desarrolladores a identificar, comprender y mitigar los sesgos en los modelos de IA (Bellamy, y otros, 2019). Además, IBM ha implementado políticas de inclusión para asegurar que sus equipos de desarrollo de IA incluyan una representación diversa, lo que ha demostrado ser un factor clave para reducir sesgos en los algoritmos. |
| Crear canales de comunicación donde los usuarios puedan reportar posibles sesgos de género en los resultados, permitiendo ajustes basados en el feedback. Enfoque: Identificación y mitigación de sesgos de género. | Google Translate permite a los usuarios reportar errores o posibles sesgos, incluido el sesgo de género, a través de un formulario de comentarios disponible en la parte inferior de la página. Para hacerlo, los usuarios deben acceder a Google Translate, ingresar el texto a traducir, y luego desplazarse hasta la opción “Enviar comentarios” al final de la interfaz. Al hacer clic en esta opción, se abre un formulario donde se pueden detallar los problemas identificados, como el sesgo de género, proporcionando ejemplos específicos (Google Translate, s.f.). |

Fuente: Consultoría BetaGroup



La etapa de explotación y supervisión es clave para mantener la equidad y responsabilidad de los sistemas de IA a lo largo del tiempo. Las buenas prácticas incluyen monitoreo constante, gobernanza ética y documentación de aprendizajes, permitiendo identificar y mitigar sesgos de género, promover mejoras continuas y garantizar una operación justa y transparente.

Tabla 7. Buenas prácticas en la etapa de explotación y supervisión (Monitoreo y ajuste)

| Descripción / Enfoque | Ejemplo de Aplicación |
|--|---|
| <p>Realizar evaluaciones regulares del impacto de género de los sistemas de IA en uso, revisando cómo afectan a las poblaciones subrepresentadas y ajustando las prácticas según los hallazgos.</p> <p>Enfoque: Identificación y mitigación de sesgos de género.</p> | <p>El Grupo de Trabajo de Sistemas de Decisión Automatizados de la Ciudad de Nueva York, establecido por la Ley Local 49 de 2018, tiene como objetivo revisar y recomendar procesos para garantizar que los sistemas de decisión automatizados utilizados por la ciudad sean justos y equitativos. Su labor incluye la realización de evaluaciones periódicas para identificar y mitigar posibles sesgos en los algoritmos empleados, promoviendo así la transparencia y la rendición de cuentas en el uso de la inteligencia artificial en el sector público (Derechos Digitales, 2019).</p> |
| <p>Establecer un enfoque de mejora continua con roles, métricas e instancias de evaluación y retroalimentación.</p> <p>Enfoque: Identificación y mitigación de sesgos de género.</p> | <p>IBM ha implementado un enfoque integral para la gobernanza de la inteligencia artificial (IA), estableciendo controles organizacionales a través de su Comité de Ética de IA y un Programa de Gobierno Integrado. Estos mecanismos aseguran que el desarrollo y despliegue de sus sistemas de IA se realicen de manera ética y responsable. Además, IBM ofrece soluciones como watsonx.governance para ayudar a otras organizaciones a implementar controles organizacionales y técnicos en sus sistemas de IA (World Economic Forum, 2024).</p> |
| <p>Documentar las lecciones aprendidas en cada fase del ciclo de vida del sistema de IA para promover una mejora continua y facilitar la replicación de buenas prácticas en otros proyectos de IA.</p> <p>Enfoque: Identificación y mitigación de sesgos de género.</p> | <p>La recomendación sobre la ética de la inteligencia artificial de la UNESCO establece la importancia de documentar las lecciones aprendidas en cada fase del ciclo de vida del sistema de IA. Esta documentación permite promover una mejora continua en el desarrollo de IA y facilita la replicación de buenas prácticas en otros proyectos de IA (UNESCO, 2024).</p> |

Fuente: Consultoría BetaGroup



La etapa de retirada o desmantelamiento de sistemas de IA es crucial para gestionar de manera ética los impactos de sesgos irreparables, evitando la perpetuación de desigualdades. Las buenas prácticas incluyen protocolos claros para minimizar daños, proteger a los afectados y reforzar la responsabilidad social, garantizando confianza en el desarrollo de tecnologías futuras.

Tabla 8. Buenas prácticas en la etapa de retirada/desmantelamiento

| Descripción / Enfoque | Ejemplo de Aplicación |
|---|--|
| En caso de que un sistema de IA presente sesgos irreparables, tener un protocolo de desmantelamiento y un plan de retiro que minimice el daño y evite la perpetuación de dichos sesgos. Enfoque: Mitigación de sesgos de género. | Un ejemplo notable de una empresa que implementó un protocolo de desmantelamiento para minimizar el daño y evitar la perpetuación de sesgos es Amazon. En 2018, la compañía desarrolló una herramienta de inteligencia artificial para la selección de personal que, tras su implementación, mostró sesgos de género al favorecer a candidatos masculinos. Al identificar este problema, Amazon decidió discontinuar el uso de dicha herramienta para evitar decisiones de contratación injustas y sesgadas (The Verge, 2018). |

Fuente: Consultoría BetaGroup

Las buenas prácticas en el ciclo de vida de los sistemas de inteligencia artificial (IA) son esenciales para identificar, prevenir y mitigar los sesgos de género. Estas incluyen la evaluación temprana de sesgos, la transparencia en los modelos y la validación constante de datos, con el objetivo de desarrollar IA inclusiva y equitativa. Al implementar estas acciones, se pueden corregir desigualdades desde las etapas iniciales, promoviendo un diseño ético y responsable. El monitoreo continuo y la participación de equipos diversos garantizan que los sistemas se adapten a los cambios sociales y las necesidades de las comunidades afectadas, fomentando un entorno digital más justo.

Asimismo, la equidad de género en la IA requiere que las herramientas y metodologías estén disponibles y sean conocidas ampliamente. Identificar los riesgos desde etapas tempranas y establecer mecanismos de mitigación debe ser una prioridad en los proyectos de IA, especialmente en el sector público. Auditar los datos y ajustar los modelos de manera regular son acciones clave para evitar la perpetuación de estereotipos de género, mientras que la retroalimentación constante fortalece los principios de justicia e inclusión en estas tecnologías.



4

Buenas Prácticas para la Adopción de los Sistemas de IA

La adopción de sistemas de inteligencia artificial exige un enfoque estructurado que permita identificar y mitigar riesgos desde sus primeras etapas. Esto implica incorporar prácticas como la evaluación de impacto ético, el diseño de sistemas explicables y la implementación de mecanismos claros para proteger la privacidad y garantizar la diversidad en los datos utilizados. Además, es fundamental establecer canales efectivos de participación y retroalimentación que permitan a las partes interesadas identificar y corregir posibles sesgos, asegurando que las decisiones tecnológicas reflejen un compromiso con la equidad y la responsabilidad social.

Buenas Prácticas para el Adopción de los Sistemas de IA

Tabla 9. Buenas prácticas para la adopción de los sistemas de IA

| Descripción / Enfoque | Ejemplo de Aplicación |
|--|---|
| <p>Realizar un análisis exhaustivo de los riesgos éticos, sociales y técnicos, con especial atención a los potenciales sesgos y discriminación que el sistema de IA pueda introducir.</p> <p>Enfoque: Identificación de sesgos de género.</p> | <p>Un ejemplo destacado de esta práctica es el Ethical AI Risk Assessment implementado por Google AI. Google realiza evaluaciones de riesgos éticos y sociales al evaluar los posibles sesgos de género y raza en sus sistemas de IA, a través de auditorías internas y la colaboración con expertos externos para identificar riesgos de discriminación en la adopción de sus sistemas (Google, 2023).</p> |
| <p>Llevar a cabo un análisis de impacto para anticipar los efectos del sistema de IA sobre los diferentes grupos sociales, considerando aspectos como equidad de género, no discriminación y otros factores éticos.</p> <p>Enfoque: Identificación de sesgos de género.</p> | <p>IBM Watson lleva a cabo estudios de impacto ético en sus sistemas de IA, involucrando a diversos grupos sociales y evaluando los efectos que sus productos pueden tener sobre las comunidades de género y minorías. Esta práctica incluye un análisis continuo de la equidad de género, asegurando que los algoritmos no reproduzcan estereotipos discriminatorios (IBM, 2023).</p> |
| <p>Verificar que el sistema de IA mantenga la explicabilidad, de tal manera que permita comprender cómo y por qué se generan los resultados.</p> <p>Enfoque: Prevención de sesgos de género.</p> | <p>Microsoft ha implementado “Fairlearn”, una herramienta que permite a los desarrolladores y usuarios de IA comprender cómo sus modelos toman decisiones, ayudando a identificar cualquier sesgo de género en los resultados y facilitando ajustes para mejorar la transparencia y explicabilidad (Microsoft, 2020).</p> |
| <p>Verificar que la adopción de IA cumpla con las leyes de protección de datos y privacidad, mediante prácticas de anonimización y cifrado, reduciendo el riesgo de exposición indebida.</p> <p>Enfoque: Prevención de sesgos de género.</p> | <p>Apple ha implementado privacidad diferencial en sus sistemas de IA, como Siri, para garantizar la anonimización de los datos de los usuarios, evitando que se recojan datos sensibles relacionados con el género o cualquier otra característica personal, cumpliendo con la normativa de protección de datos como el GDPR (Apple, 2019).</p> |
| <p>Revisar que existan canales de comunicación para que los usuarios puedan expresar preocupaciones o sugerencias, y recibir respuestas sobre cómo se abordarán sus comentarios.</p> <p>Enfoque: Mitigación de sesgos de género.</p> | <p>Google permite a los usuarios reportar problemas relacionados con el sesgo de género a través de la opción “Enviar comentarios” en Google Translate. Este mecanismo de retroalimentación garantiza que las preocupaciones sobre posibles sesgos sean atendidas, ajustando las traducciones y mejorando la inclusión en el servicio (Google, 2020).</p> |

Fuente: Consultoría BetaGroup

La implementación de sistemas de inteligencia artificial representa una oportunidad significativa para avanzar hacia sociedades más justas e inclusivas. Sin embargo, esto requiere un compromiso continuo con la actualización y mejora de las prácticas adoptadas, adaptándolas a los cambios tecnológicos, normativos y sociales. Fomentar una cultura de responsabilidad compartida entre desarrolladores, reguladores, organizaciones y usuarios finales es esencial para garantizar que la IA sea una herramienta que no solo evite la perpetuación de desigualdades, sino que también promueva activamente la equidad y el bienestar de todas las personas.

5

Catálogo de Herramientas y Árbol de Decisión

Esta sección presenta dos recursos fundamentales para apoyar el desarrollo, implementación y supervisión de sistemas de inteligencia artificial (IA) inclusivos, diseñados específicamente para abordar y mitigar los sesgos de género. Estos recursos están dirigidos a diversas personas o entidades clave, como los equipos de desarrollo de IA, responsables de crear tecnologías inclusivas desde las etapas iniciales; grupos de defensa de género, que aseguran la representatividad y ética en los datos utilizados; y académicos e investigadores, que avanza en la innovación técnica con un enfoque inclusivo. También están orientados a empresas tecnológicas, encargadas de implementar soluciones éticas, reguladores y organismos gubernamentales, que supervisan la alineación de la IA con políticas públicas, y personas usuarias finales, quienes son centrales en este proceso, al ser protegidas de impactos negativos y participar en la mejora continua de estas tecnologías. Además, los medios de comunicación y periodistas tecnológicos son considerados actores clave en la sensibilización y comunicación de los impactos sociales de la IA.

Ambos recursos están diseñados para fomentar un ecosistema tecnológico ético e inclusivo, proporcionando herramientas prácticas y accesibles que contribuyan al desarrollo responsable y equitativo de la inteligencia artificial. El DNP hará disponible este catálogo mediante una plataforma en línea con una interfaz intuitiva, accesible desde cualquier navegador web y dispositivo con conexión a Internet, sin necesidad de instalar software adicional. Esta plataforma garantizará que diversas personas o entidades clave puedan acceder fácilmente a las herramientas y metodologías necesarias para identificar, prevenir y mitigar sesgos de género en los sistemas de IA. Este enfoque asegura un acceso equitativo, fomentando la adopción de soluciones inclusivas y fortaleciendo la equidad en el desarrollo y la implementación de la IA.

5.1

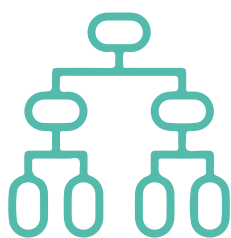
Catálogo de Herramientas



El Catálogo de Herramientas reúne recursos prácticos como software, librerías, frameworks, metodologías, listas de verificación y casos de uso, diseñados para identificar, prevenir y mitigar sesgos de género en la IA. Estas herramientas son de acceso libre, factibles de implementar en contextos locales, y están orientadas a promover la equidad de género y la creación de sistemas inclusivos. Además, el catálogo proporciona guías técnicas y ejemplos aplicados que facilitan el diseño y la evaluación responsable de soluciones tecnológicas.

5.2

Árbol de Decisión



El Árbol de Decisión es una herramienta estructurada diseñada para guiar a las personas usuarias en la identificación y abordaje de posibles sesgos de género en cada etapa del desarrollo, adquisición o implementación de sistemas de inteligencia artificial (IA). Facilita la toma de decisiones informadas al ofrecer criterios claros sobre datos, metodologías y herramientas, además de proponer acciones específicas para mitigar los riesgos detectados. Su enfoque interactivo garantiza la incorporación de principios éticos y de equidad en todas las fases del ciclo de vida de la IA, promoviendo soluciones responsables e inclusivas.



6

Recomendaciones para la Actualización y Mantenimiento de la Guía de Buenas Prácticas



Establecer revisiones periódicas de la guía para asegurar que las prácticas recomendadas reflejen los avances tecnológicos y las normativas actualizadas sobre ética y equidad en IA.



Actualizar la guía conforme a nuevas regulaciones, estándares éticos internacionales y buenas prácticas emergentes que refuercen la identificación, prevención y mitigación de sesgos de género.



Recopilar retroalimentación de los usuarios de la guía y documentar las lecciones aprendidas de implementaciones previas para ajustar y mejorar las prácticas recomendadas.



Evaluar cómo los cambios sociales y culturales pueden influir en los criterios de equidad de género, ajustando las prácticas para que sigan siendo inclusivas y relevantes en diferentes contextos.



Generar procesos formativos y de capacitación para los responsables de la actualización de la guía en temas relacionados con: Normativas y regulación en IA con enfoque de género, principios éticos en IA, sesgos de género en IA, identificación de sesgos en los datos, métodos y herramientas de auditoría de sesgos, técnicas de mitigación y ajuste de modelos, diseño de procesos inclusivos, evaluación de impacto social y monitoreo continuo, entre otras, con el fin de mantener una comprensión actualizada de los sesgos de género y los principios éticos en IA.



Implementar herramientas de evaluación de impacto que midan la efectividad de la guía en la identificación, prevención y mitigación de sesgos de género, ajustando las prácticas según los resultados obtenidos.

7 Bibliografía

Accenture. (s.f.). Trabajar para acelerar la igualdad para todos. Obtenido de <https://www.accenture.com/co-es/about/inclusion-diversity/gender-equality#:~:text=Hemos%20establecido%20metas%20ambiciosas%20para,para%20personas%20con%20g%C3%A9nero%20binario>.

AI4ALL. (2020). Building the Future of AI: Diverse Talent, Inclusive AI Education. Obtenido de <https://ai-4-all.org/>

Apple. (28 de Agosto de 2019). Mejoras en las protecciones de privacidad de Siri. Obtenido de <https://www.apple.com/co/newsroom/2019/08/improving-siris-privacy-protections/>

Apple. (s.f.). Privacy. That's Apple. Obtenido de <https://www.apple.com/privacy/>

Bellamy, R., Dey, K., Hind, M., Hoffman, S., Houde, S., Kannan, K., . . . Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. IBM Journal of Research and Development, 63(4/5), 63(4/5), 4-1.

Buolamwini, J., & Gebru, T. (2018). When women or minorities are underrepresented in datasets, algorithms are likely to perform poorly for these groups, reinforcing existing inequalities. Proceedings of Machine Learning Research, (págs. 1-15).

CONPES. (18 de Abril de 2022). Política Pública de Equidad de Género para las Mujeres: Hacia el Desarrollo Sostenible del País. Obtenido de <https://colaboracion.dnp.gov.co/cdt/Conpes/Econ%C3%B3micos/4080.pdf>

Derechos Digitales. (2019). Políticas públicas e Inteligencia Artificial. Una lista de buenas prácticas para tener en consideración. Obtenido de https://ia.derechosdigitales.org/wp-content/uploads/2022/05/DD_IA_03.pdf

Exponiendo.ai. (2019). La diversidad en los rostros de IBM. Obtenido de https://exposing.ai/ibm_dif/

Friedman, B., & Nissenbaum, H. (1996). Bias in Computer Systems. ACM Transactions on Information Systems, 330-347.

GLAAD. (24 de Junio de 2021). GLAAD se asocia con Google para ofrecer recursos de marketing LGBTQ "All In". Obtenido de <https://glaad.org/announcing-glaad-partnership-google-all/>

Google. (6 de Diciembre de 2018). Cómo reducir el sesgo de género en Google Translate. Obtenido de <https://blog.google/products/translate/reducing-gender-bias-google-translate/>

Google. (17 de Marzo de 2020). Ayudas de Google Translate. Obtenido de https://support.google.com/translate/answer/12111375?hl=es-ES&utm_source=chatgpt.com

Google. (Octubre de 2023). Evaluating social and ethical risks from generative AI. Obtenido de https://deepmind.google/discover/blog/evaluating-social-and-ethical-risks-from-generative-ai/?utm_source=chatgpt.com

Google AI. (2020). Building fairness into AI. Obtenido de <https://ai.google/research>

Google. (s.f.). Conceptos del AA. . Obtenido de <https://developers.google.com/machine-learning/crash-course/fairness?hl=es-419>

Google. (s.f.). What-If Tool. Obtenido de <https://pair-code.github.io/what-if-tool/>

Google Translate. (s.f.). Google Translate. Obtenido de <https://translate.google.com/?hl=es&sl=en&tl=es&op=translate>

IBM. (2023). Informe IBM Impact de 2023. Obtenido de https://www.ibm.com/es-es/impact/2023-ibm-impact-report?utm_source=chatgpt.com

IBM. (25 de Octubre de 2024). Evaluaciones de equidad. Obtenido de <https://dataplatform.cloud.ibm.com/docs/content/wsj/model/wos-monitor-fairness.html?context=cpdaas&locale=es>

IBM. (s.f.). AI Fairness 360. Obtenido de <https://aif360.res.ibm.com/>

League, A. J. (s.f.). Obtenido de https://www-ajl-org.translate.google/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es-419&_x_tr_pto=sc

Microsoft. (19 de Mayo de 2020). AI The Show. Obtenido de <https://learn.microsoft.com/es-es/shows/ai-show/building-fairer-ai-systems-with-fairlearn>

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

OCDE. (2 de Mayo de 2024). Recommendation of the Council on Artificial Intelligence. Obtenido de <https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449>

Raji, I., & Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (págs. 429-435). New York: Association for Computing Machinery.

The Verge. (10 de Octubre de 2018). Amazon descarta una herramienta de reclutamiento de inteligencia artificial interna que estaba sesgada contra las mujeres. Obtenido de <https://www.theverge.com/2018/10/10/17958784/ai-recruiting-tool-bias-amazon-report>

UNESCO. (Noviembre de 2021). Recomendación sobre la ética de la inteligencia artificial. Obtenido de <https://unesdoc.unesco.org/ark:/48223/pf0000380455.page>

UNESCO. (2024). Challenging Systematic Prejudices: An Investigation into Bias Against Women and Girls in Large Language Models. París: UNESCO.

UNESCO. (Febrero de 2024). Ética de la inteligencia artificial. La recomendación. Obtenido de <https://www.unesco.org/es/artificial-intelligence/recommendation-ethics>

UNFPA. (s.f.). Preventing Technology-Facilitated Gender-Based Violence (TF GBV). Obtenido de https://www.un.org/techenvoy/sites/www.un.org.techenvoy/files/GDC-Submission_UNFPA.pdf

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., & Schwartz, O. (6 de Diciembre de 2018). AI now report 2018. Obtenido de <https://ainowinstitute.org/publication/ai-now-2018-report-2>

World Economic Forum. (20 de Septiembre de 2024). Gobernanza de la IA: Cómo la regulación, la colaboración y la demanda de habilidades están configurando las tendencias. Obtenido de <https://es.weforum.org/stories/2024/09/gobernanza-de-la-ia-como-la-regulacion-la-colaboracion-y-la-demanda-de-talento-estan-dando-forma-a-la-industria/>

Ziwei, L., Ping, L., Xiaogang, W., & Xiaoou, T. (24 de Septiembre de 2015). Large-scale CelebFaces Attributes (CelebA) Dataset. Deep Learning Face Attributes in the Wild. Obtenido de <https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>



Beta Group