INFORME EJECUTIVO: OPORTUNIDADES, RIESGOS Y LIMITACIONES DE LA INTELIGENCIA ARTIFICIAL

Análisis de Sesgos Algorítmicos, Inteligencia Artificial Responsable (Responsible AI), IA Generativa y Gobernanza

Organizador: Asociación REDEL (Red de Entidades de Desarrollo Local) | Proyecto: Copilotos (Bizkaia)

Cofinancia: Diputación Foral de Bizkaia | Ponente: Nerea Luis Mingueza (Doctora en Inteligencia Artificial, Consultora y Divulgadora)



1. SÍNTESIS EJECUTIVA Y ENCUADRE INSTITUCIONAL

El webinar «Oportunidades, Riesgos y Limitaciones de la Inteligencia Artificial», organizado por la Asociación REDEL (Red de Entidades de Desarrollo Local), se enmarca operativamente en el Proyecto Copilotos, una iniciativa desarrollada en el territorio histórico de Bizkaia cofinanciada por la Diputación Foral de Bizkaia. El objetivo estratégico de Copilotos consiste en capacitar e integrar herramientas de Inteligencia Artificial en los itinerarios y procesos de empleo y emprendimiento local, estructurando programas formativos específicos para personas desempleadas, emprendedores y el personal técnico municipal de orientación e inserción laboral.

Impartida por la doctora en Inteligencia Artificial Nerea Luis Mingueza —con una dilatada trayectoria que aúna la investigación académica, la ingeniería en la industria y la consultoría estratégica—, la sesión trasciende el discurso habitual centrado exclusivamente en el potencial tecnológico para sumergirse en la cara menos visible de la disciplina: los fallos estructurales, las crisis operativas, los sesgos algorítmicos y las limitaciones materiales que condicionan su adopción real. La tesis central sostiene que la Inteligencia Artificial es una tecnología extraordinaria pero imperfecta, cuyo despliegue sostenible requiere abandonar la fascinación acrítica y asumir una rigurosa perspectiva de IA Responsable (Responsible AI).

2. GENEALOGÍA DE LOS SESGOS ALGORÍTMICOS Y EL DILEMA DE LA CAJA NEGRA

Durante las primeras etapas del desarrollo del aprendizaje automático (Machine Learning), los modelos se diseñaban para optimizar de forma aislada métricas matemáticas de precisión y acierto en tareas muy acotadas (como el cribado radiológico o la detección de defectos industriales). Sin embargo, el punto de inflexión se produjo entre 2017 y 2018, cuando los algoritmos comenzaron a desplegarse a escala comercial sobre seres humanos, específicamente a través del reconocimiento facial y la clasificación biométrica. En ese momento se constató que una alta precisión media no garantiza equidad, sino que con frecuencia oculta profundas asimetrías de rendimiento en perjuicio de colectivos minoritarios o vulnerables.

Los algoritmos de aprendizaje profundo operan en gran medida como «cajas negras»: su inmensa complejidad matemática y sus millones de parámetros impiden trazar de principio a fin la cadena causal que lleva de una entrada de datos a una decisión final. El sesgo algorítmico surge fundamentalmente de tres fallas metodológicas en el ciclo de vida del dato:

- **Sesgo en la selección del dato:** Desequilibrios estructurales en los conjuntos de datos históricos de entrenamiento, donde ciertos colectivos están infrarrepresentados o ausentes.
- **Sesgo en el etiquetado:** Prejuicios y asociaciones culturales implícitas proyectados por los equipos humanos encargados de clasificar y etiquetar las muestras.
- **Sesgo de optimización:** Fijación ciega de patrones estadísticos y estereotipos consolidados que el modelo asume como verdades predictivas inmutables.

3. CASOS DE ESTUDIO EMBLEMÁTICOS DE DISCRIMINACIÓN Y AUDITORÍA

La ponente desgranó varios casos de repercusión mundial que motivaron la intervención de reguladores y el nacimiento de auditorías algorítmicas independientes:

- **Algoritmo COMPAS (Justicia Penal en EE. UU.):** Sistema predictivo utilizado por el sistema judicial y policial en Estados Unidos para predecir la probabilidad de reincidencia de personas excarceladas mediante una escala del 0 al 10. Una auditoría externa llevada a cabo por la organización periodística ProPublica demostró que, ante perfiles delictivos y circunstancias comparables, el algoritmo asignaba sistemáticamente puntuaciones de riesgo significativamente más altas a personas negras que a personas blancas.
- **Estudio 'Gender Shades' y el Documental 'Coded Bias':** Investigación pionera liderada por Joy Buolamwini y Timnit Gebru que evaluó los sistemas comerciales de reconocimiento facial de multinacionales como IBM, Microsoft y Face++. El estudio demostró que mientras los sistemas alcanzaban precisiones cercanas al 99% en hombres de tez clara, el índice de error se disparaba drásticamente al evaluar a mujeres de tez oscura (con tasas de error de hasta el 35%). El documental 'Coded Bias' inmortalizó cómo Buolamwini tuvo que confeccionar una máscara blanca en el MIT Media Lab para que el software del laboratorio detectara su rostro como humano.
- **Algoritmo PULSE y Despersonalización Étnica:** Algoritmos de reconstrucción de imágenes que, al intentar desempaquetar y renderizar rostros pixelados de figuras públicas como el presidente Barack Obama o la congresista Alexandria Ocasio-Cortez, transformaban sus rasgos convirtiéndolos por defecto en personas blancas de origen caucásico, reflejando el sesgo homogeneizador de sus bases de datos.
- **Auditoría de Reconocimiento de Objetos (Vicomtech / Euskadi):** Experimento desarrollado por la empresa vasca Vicomtech que procesó fotografías de hombres y mujeres en herramientas estándar de reconocimiento de objetos (como las de Amazon). Al introducir la misma imagen modificando únicamente el género del sujeto, el reconocedor asociaba al varón con herramientas mecánicas y taladros de potencia, mientras que en la mujer identificaba productos y tareas de limpieza con más de un 82% de probabilidad, suprimiendo la detección de herramientas.

- **Sesgos en Procesos de Selección de Personal:** Plataformas automatizadas de filtrado de currículums que penalizan candidaturas femeninas en puestos de perfil técnico o directivo al asociar el éxito laboral histórico con palabras y trayectorias masculinas. Por este motivo, el Reglamento Europeo de Inteligencia Artificial (AI Act) ha tipificado expresamente estos sistemas como de 'Alto Riesgo', sometiéndolos a estrictas auditorías de equidad y no discriminación.
- **Sesgos en Robótica Social:** Investigaciones de la profesora Ayanna Howard demostraron que la inmensa mayoría de robots de asistencia social son diseñados de color blanco. Experimentos comparativos con réplicas idénticas en color negro generaron reacciones de desconcierto y atribución de hostilidad en observadores humanos, demostrando cómo los estereotipos sociales se trasladan inadvertidamente al hardware.

4. PILARES DE LA INTELIGENCIA ARTIFICIAL RESPONSABLE (RESPONSIBLE AI)

Frente al concepto de «AI Safety» (que parte de la premisa implícita de que la tecnología es insegura por naturaleza), la comunidad científica e industrial ha consolidado el paradigma de la Inteligencia Artificial Responsable (Responsible AI), estructurado en tres grandes disciplinas técnicas y metodológicas:

Explicabilidad (XAI - Explainable AI)

Desarrollo de herramientas matemáticas capaces de abrir la 'caja negra' para auditar cómo y por qué se toma una decisión. En datos tabulares (banca, seguros y concesión de crédito), destacan los 'SHAP values' (Shapley Additive Explanations), que ponderan de forma transparente qué variables individuales pesan más en la resolución. En visión por computador, se emplean 'mapas de calor' (Saliency Maps e Integrated Gradients) para verificar si el algoritmo detecta la patología médica real o si se distrae con artefactos espurios (como ocurrió en modelos de detección de fracturas que aprendieron a identificar las marcas dejadas con rotulador por los radiólogos en lugar de la rotura ósea).

Transparencia y Hojas de Modelo (Model Cards)

Iniciativa estandarizada por corporaciones como Google, Microsoft y OpenAI que funciona como un 'prospecto farmacéutico' para el software de IA. Detalla la arquitectura del modelo, datos de

entrenamiento, limitaciones declaradas, métricas de justicia distributiva (fairness) y casos de uso expresamente contraindicados para prevenir daños reputacionales y legales.

Responsabilidad, Rendición de Cuentas (Accountability) y Marco Ético

Establecimiento de fronteras legales infranqueables. En la Unión Europea, el Reglamento de IA prohíbe taxativamente prácticas como la puntuación social ('social scoring') y la videovigilancia biométrica masiva en espacios públicos. Asimismo, se ilustra la imposibilidad de codificar una 'ética universal': el experimento del MIT 'Moral Machine' demostró que mientras la cultura occidental prioriza salvar a niños en dilemas de vehículos autónomos, la cultura oriental prioriza salvar a los ancianos.

5. LA ERA DE LA IA GENERATIVA: ALUCINACIONES, AGENTES Y RESPONSABILIDAD LEGAL

La irrupción de modelos fundacionales y herramientas conversacionales (como ChatGPT) supuso un salto de escala cualitativo: la IA pasó de clasificar datos a crear contenidos inéditos (texto, audio, imagen y código). Sin embargo, esta capacidad creativa amplifica exponencialmente los riesgos operativos:

- **Estereotipos en Modelos Generativos de Imagen:** Las primeras versiones comerciales presentaban una marcada tendencia a asociar conceptos cotidianos con estereotipos hegemónicos (por ejemplo, generar exclusivamente parejas heterosexuales y caucásicas ante la instrucción de una boda).
- **Alucinaciones y el Salto a los Modelos Razonadores:** Respuestas convincentes pero completamente falsas cuando el modelo carece de información veraz. La ponente destacó la evolución hacia modelos con capacidades de 'razonamiento' (como las arquitecturas contemporáneas de OpenAI), que autoevalúan de forma iterativa sus respuestas internas antes de emitir un veredicto definitivo, mitigando las alucinaciones.
- **El Precedente Jurídico de Air Canada:** Sentencia judicial histórica contra la aerolínea Air Canada. La compañía fue condenada a indemnizar a un pasajero tras aplicarle una tarifa errónea recomendada por el chatbot de su página web. El tribunal sentenció que el asistente virtual constituye un canal oficial de comunicación corporativa, siendo la empresa 100% responsable de las declaraciones, acuerdos o errores emitidos por su algoritmo.

- **Clonación de Voz, Deepfakes y Avatares Digitales:** Aparición de avatares sintéticos fotorrealistas con sincronización labial avanzada (como la campaña publicitaria de Trivago), acuerdos de cesión de imagen de celebridades (Bruce Willis o Snoop Dogg) frente a demandas por clonación no consentida de voz (Bad Bunny), y la proliferación de fraudes mediante 'vishing' (audios breves de WhatsApp suplantando a directivos o familiares).
- **El Trabajo Oculto del Etiquetado Humano:** Persistencia de cientos de miles de trabajadores en países en desarrollo sometidos a salarios precarios y jornadas extenuantes para etiquetar datos y filtrar contenidos violentos o tóxicos para las grandes tecnológicas.

6. SOSTENIBILIDAD MEDIOAMBIENTAL, GEOPOLÍTICA Y ENTORNO REGULATORIO

La escalabilidad de la Inteligencia Artificial colisiona directamente con los límites físicos y energéticos del planeta. El entrenamiento e inferencia masiva de modelos exige centros de datos que demandan un consumo exponencial de electricidad y agua de refrigeración. La sostenibilidad se ha convertido en una variable geoestratégica de primer orden:

- **Estados Unidos:** Mega-inversiones en infraestructura de supercomputación soberana (iniciativa Stargate).
- **Unión Europea:** Inversión en factorías de IA ('AI Factories'), una de las cuales se proyecta en España.
- **China:** Estrategia combinada de grandes infraestructuras e investigación puntera en modelos abiertos y de bajo coste energético.

Para conjugar innovación y seguridad jurídica, España ha sido pionera con el lanzamiento del primer 'Sandbox' regulatorio de Inteligencia Artificial de la Unión Europea, permitiendo a las empresas testear sus algoritmos en un entorno controlado para evaluar su conformidad con el Reglamento Europeo de IA (AI Act), cuya plena aplicación se desplegará de cara a 2027.

7. MATRIZ DE RIESGOS, MANIFESTACIONES Y MEDIDAS DE MITIGACIÓN

Categoría de Riesgo	Manifestaciones y Casos Prácticos	Medidas de Mitigación y Gobernanza
Sesgos y Discriminación Algorítmica	COMPAS (reincidencia sesgada por raza), Gender Shades (fallos en reconocimiento de mujeres negras), PULSE y cribado de currículums en RRHH.	Auditorías algorítmicas independientes, desanonimización de métricas de rendimiento y balanceo riguroso de datos de entrenamiento.
Falta de Explicabilidad ('Caja Negra')	Decisiones automatizadas en denegación de créditos bancarios o diagnósticos médicos basados en marcas accidentales en radiografías.	Implementación de técnicas XAI (SHAP values para datos tabulares y mapas de calor/saliency maps para visión por computador).
Alucinaciones y Errores Operativos	Chatbots comerciales que inventan políticas de empresa o información falsa (caso judicial Air Canada).	Barreras de protección (guardrails), modelos con capacidades de razonamiento reflexivo e integración de verificación humana (human-in-the-loop).
Fraude Digital e Inseguridad	Estafas de suplantación de identidad mediante clonación de voz (vishing en WhatsApp) y generación no consentida de deepfakes.	Verificación de canal por doble factor, formación ciudadana en pensamiento crítico y marcos de ciberseguridad especializada (OWASP para LLMs).
Huella Ecológica y Límites Físicos	Consumo intensivo de electricidad y agua en centros de datos derivado del auge masivo de consultas.	Desarrollo de modelos destilados ('Small Language Models'), centros de datos abastecidos por renovables y optimización algorítmica.

8. CONCLUSIONES Y RECOMENDACIONES PARA EL DESARROLLO LOCAL Y EL EMPLEO

La adopción de la Inteligencia Artificial en el ámbito municipal y en las agencias de desarrollo local debe guiarse por los siguientes principios estratégicos:

1. **Supervisión Humana Permanente (Human-in-the-Loop):** En los procesos de intermediación laboral y cribado de candidatos (como en el Proyecto Copilotos de Bizkaia), la decisión final nunca debe delegarse ciegamente en un algoritmo. Los profesionales de orientación deben supervisar las propuestas algorítmicas para erradicar cualquier sesgo de género, edad u origen.
2. **Alfabetización Digital Crítica:** Comprender que la IA se integrará de forma embebida en aplicaciones de uso cotidiano (suites de ofimática, sistemas de gestión interna y diseño gráfico). Por tanto, la capacitación del personal municipal y de los usuarios debe orientarse al desarrollo del pensamiento crítico, la verificación de fuentes y el dominio de prompts precisos.
3. **Responsabilidad Institucional en Canales de Atención:** Como demostró el caso de Air Canada, cualquier asistente conversacional implementado en portales municipales o agencias públicas compromete legal e institucionalmente a la entidad. Es preceptivo incorporar salvaguardas que reconozcan las limitaciones de respuesta y deriven a operadores humanos ante consultas complejas o normativas.
4. **Transparencia Contractual y Cumplimiento Normativo:** Exigir a los proveedores tecnológicos la entrega de Model Cards que especifiquen los datos de entrenamiento y las pruebas de sesgo efectuadas, preparándose para las exigencias de cumplimiento del Reglamento Europeo de Inteligencia Artificial.

MENSAJE CLAVE PARA AGENCIAS Y ADMINISTRACIONES LOCALES

Conclusión de Cierre:

La Inteligencia Artificial no reemplazará a los profesionales del desarrollo local, pero los profesionales y las organizaciones que comprendan sus limitaciones, gestionen sus sesgos y utilicen estas herramientas con criterio ético liderarán una transformación pública mucho más eficaz, justa y cercana a la ciudadanía.



[Enlace a la grabación]