unesco

Kit de herramientas global sobre IA y el Estado de derecho para el poder judicial

Publicado en 2023 por
La Organización de las Naciones Unidas
para la Educación, la Ciencia y la Cultura
7, place de Fontenoy, 75352 París 07 SP, Francia

ISBN

Este kit de herramientas fue preparado por:

Dra. Miriam Stankovich, Especialista principal en política digital del Center for Digital Acceleration (Bethesda, Maryland, Estados Unidos).

La sección sobre sesgo de *IA e igualdad de género* fue desarrollada por Ivana Feldfeber (Cofundadora y Directora ejecutiva de DataGénero), Yasmín Quiroga (Cofundadora de DataGénero y Secretaria en el Tribunal Penal N.º 10 de Buenos Aires, Argentina) y Marianela Ciolfi Felice (Profesora asistente en Diseño de interacción en la Universidad KTH, Suecia, y Asesora en DataGénero). La sección sobre *Oportunidades: IA y el Poder Judicial en el Continente Africano* fue escrita por el Prof. Vukosi Marivate (Universidad de Pretoria, Sudáfrica).

Asesores académicos:

Prof. Joan Barata Mir (Profesor emérito en Justitia, Dinamarca-Estados Unidos), Prof. Maria Fasli (Universidad de Essex, Reino Unido), Prof.ª Els de Busser (Universidad de Leiden, Países Bajos) y Prof. Vukosi Marivate (Universidad de Pretoria, Sudáfrica).

Revisores UNESCO:

Cedric Wachholz, Jaco Du Toit, Bhanu Neupane, Rosa María González, Natalia Zuazo, Misako Ito, Mehdi Benchelah.

Revisores externos:

Jhalak M. Kakkar (Director ejecutivo, Centro para la Gobernanza de la Comunicación, Universidad Nacional de Derecho de Delhi y Profesor visitante, NLU Delhi), Nidhi Singh (Oficial de Programas, Centro para la Gobernanza de la Comunicación, NLU Delhi), Juez Jean Aloise Ndiaye (Corte Suprema de Senegal), Dra. Alexandre Barbosa (Jefe del Centro Regional de Estudios sobre el Desarrollo de la Sociedad de la Información, Cetic.br | NIC.br), Luiz Costa (Observatorio Brasileño de Inteligencia Artificial, OBIA), Ameen Jauhar (Jefe del equipo, ALTR, Vidhi Centre for Legal Policy), Nathalie Smuha (Profesora asistente de la Facultad de Derecho de KU Leuven y Emile Noël Becaria de la Facultad de Derecho de la Universidad de Nueva York), Lee Tiedrich (Miembro Distinguida de la Facultad, Ethical Tech de la Universidad de Duke y experta en IA de la OCDE), Marc Rotenberg (Presidente y fundador del Centro de IA y Política Digital), Alfonso Peralta Gutiérrez (Juez de Primera Instancia e Investigación Criminal, Granada, España), Murali Sagi (Director Ejecutivo Adjunto de la Comisión Judicial de NSW, Nueva Gales del Sur), Anthony Wong (Presidente de IFIP, Federación Internacional para el Procesamiento de la Información), Saurabh Karn (Fundador y Científico principal de OpenNyAI y Fundador de Sampatti Card), y Prof. Keith R. Fisher (Miembro distinguido, National Judicial College, EE. UU.), Niki Iliadis (Directora, AI and the Rule of Law en TFS, The Future Society), Amanda Leal (Asociada, AI Governance en TFS), Nicolas Miailhe (Fundador y Presidente de TFS), Prof. Srikrishna Deva Rao (Vicerrector de la Universidad de Derecho NALSAR, Hyderabad), Sr. Pranav Verma (Profesor Asistente en la Facultad Nacional de Derecho de la Universidad de la India, Bangalore), Dr. Ravi Srinivas (Profesor Adjunto de la Universidad de Derecho NALSAR), Dr. Naveen Thayyil (Profesor Asociado en IIT, Delhi), Neela Badami (Socia en Samvad Partners), Dr. Shouvik Kumar Guha (Profesor Asociado en la Universidad Nacional de Ciencias Jurídicas de Bengala Occidental, Calcuta), Rohan George (Socio en Samvad Partners), Nehaa Chaudhari (Socia en Ikigai Law), Pallavi Sondhi (Asociada Sénior en Ikigai Law), Ajey Karthik (Asociado en Ikigai Law) y Namratha Murugeshan (Asociado en Ikigai Law), Jaideep Reddy (Abogado de Tecnología en Trilegal y Profesor Visitante en la Escuela Nacional de Derecho de la Universidad de India, Bengaluru).

Dirección y coordinación del proyecto:

Prateek Sibal, Especialista de programas, políticas digitales y transformación digital, UNESCO.
Charline d'Oultremont, Consultora, políticas digitales y transformación digital, UNESCO.
Giovanni Imperiali, Pasante, Políticas digitales y transformación digital, UNESCO.
Gustavo Fonseca Ribeiro, Consultor, Políticas digitales y transformación digital, UNESCO, contribuyó a la organización de talleres piloto para el kit de herramientas.

Traducción y corrección: Nube Consulting
Maqueta y rotulación: Nube Consulting + Alien Books
Diseño de la portada: Nube Consulting

Impreso por: UNESCO

El Kit de herramientas global se ha desarrollado como parte del proyecto financiado por la Comisión Europea "Apoyo a los Estados miembros en la aplicación de la Recomendación de la UNESCO sobre la ética de la IA a través de herramientas innovadoras"

BREVE RESUMEN

La Inteligencia artificial como nueva frontera para el poder judicial

¿Qué es la Inteligencia artificial (IA)? ¿Cómo funciona? Y lo que es más importante, ¿cómo se abre camino en el contexto judicial? Tecnologías como la IA han existido durante décadas, pero solo recientemente han comenzado a usarse en una variedad de entornos de justicia y aplicación de la ley. Si bien la IA tiene un inmenso potencial para el sistema de justicia, ya que ayuda a los jueces a tomar mejores decisiones, mejora la eficiencia, aumenta el acceso y ayuda a detectar y prevenir el delito, también existen algunas cuestiones importantes que las partes interesadas en la justicia deben tener en cuenta mientras se preparan para un futuro en el que la IA se utiliza cada vez más en los sistemas de justicia.

En 2022, la UNESCO lanzó dos evaluaciones de necesidades. En primer lugar, a través de la Encuesta de evaluación de necesidades de inteligencia artificial de la UNESCO en África, el 90 % de los 32 países encuestados solicitaron apoyo para el desarrollo de capacidades para el poder judicial en materia de IA. Al mismo tiempo, una segunda encuesta mundial de actores judiciales en 100 países subrayó la necesidad de comprender mejor el uso de la IA en la administración de justicia y sus repercusiones jurídicas más amplias en las sociedades.

El “Kit de herramientas mundial sobre la IA y el Estado de derecho” para el Poder Judicial responde a estas necesidades y proporciona a los actores judiciales (jueces, fiscales, fiscales estatales, abogados públicos, universidades de derecho e instituciones de formación judicial) el conocimiento y las herramientas necesarias para comprender los beneficios y riesgos de la IA en su trabajo. El kit de herramientas ayudará a los actores judiciales a mitigar los posibles riesgos de la IA para los derechos humanos al brindar orientación sobre las leyes, principios, normas y jurisprudencia internacional de derechos humanos relevantes que sustentan el uso ético de la IA.

unesco

“Las guerras nacen en la mente de hombres y mujeres, es en la mente de los hombres y mujeres donde deben erigirse los baluartes de la paz.”

Kit de herramientas global sobre IA y el Estado de derecho para el poder judicial

PRÓLOGO

Los jueces desempeñan un papel crucial en la protección de los derechos civiles: pueden establecer poderosos precedentes legales en sus juicios sobre casos individuales, lo que permite a un país avanzar en un área particular de la ley. Los procesos legales recientes han demostrado que el poder judicial puede recurrir al ordenamiento internacional de los derechos humanos, las salvaguardias constitucionales y las leyes de protección de datos para protegerse contra los sistemas de IA discriminatorios y sesgados. Para que los jueces desempeñen esta función primordial de manera efectiva, debemos ayudar a desarrollar su conocimiento y comprensión de cómo funcionan los sistemas de IA y cómo se puede aplicar el derecho internacional de los derechos humanos a la IA.

Desde 2014, la Iniciativa Mundial de Jueces de la UNESCO ha involucrado a más de 34 800 actores judiciales de más de 160 países en materia de libertad de expresión, acceso a la información y seguridad de los periodistas. Esta iniciativa ayuda a fortalecer las capacidades de los operadores judiciales para enfrentar los desafíos emergentes del poder judicial y proteger los derechos humanos fundamentales y la libertad de expresión.

En 2022, la Iniciativa de Jueces lanzó su programa sobre IA y el Estado de derecho con el objetivo de involucrar a las partes interesadas dentro de los Sistemas de Justicia en un debate global y oportuno sobre las aplicaciones de la inteligencia artificial y su impacto en el Estado de derecho. Lo anterior hace seguimiento a la Recomendación de la UNESCO sobre la Ética de la inteligencia artificial, un plan integral para construir regímenes regulatorios sobre valores y principios universalmente aceptados, adoptado por los 193 Estados miembros de la UNESCO en noviembre de 2021. La Recomendación subrayó el valor de los “sistemas de IA para mejorar el acceso a la información y el conocimiento” y la necesidad de “mejorar la capacidad del poder judicial para tomar decisiones relacionadas con los sistemas de IA según el Estado de derecho y en línea con el derecho y las normas internacionales”.
Tras una encuesta mundial en la que participaron actores judiciales de la Red de Antiguos Alumnos de la Iniciativa Mundial de Jueces, la UNESCO y sus socios desarrollaron un Curso masivo abierto en línea sobre IA y el Estado de derecho (MOOC) en siete idiomas en 2022. El MOOC analiza las buenas prácticas sobre cómo los tribunales deciden los casos relacionados con la IA, de acuerdo con los derechos humanos y los estándares éticos, y explora las oportunidades y los riesgos de la adopción de la IA por parte de los sistemas de justicia.

Siguiendo los pasos de este MOOC, el “Kit de herramientas global sobre IA y el Estado de derecho” tiene como objetivo capacitar a los actores judiciales sobre cómo garantizar que el desarrollo de la IA alcance su máximo potencial de acuerdo con el Estado de derecho. De hecho, mientras nos esforzamos por desarrollar nuevas leyes para gobernar la IA en sí misma, es imperativo que apoyemos a jueces, fiscales y funcionarios públicos con capacidades mejoradas para protegernos de los riesgos relacionados con la IA.

ÍNDICE

Lista de acrónimos

ACLU	Unión Americana de Libertades Civiles
ADM	Toma de decisiones algorítmica
AGR	Reconocimiento automatizado de género
AI	Inteligencia artificial
CAHAI	Comité Ad Hoc del Consejo de Europa sobre Inteligencia Artificial
ChatGPT	Transformador preentrenado generativo
COMPAS	Perfiles de gestión de delincuentes correccionales para penas alternativas
CRT	Tribunal de Resolución Civil
CE	Comisión Europea
CEDH	Convenio Europeo de Derechos Humanos
EFF	Electronic Frontier Foundation
ESI	Información almacenada electrónicamente
UE	Unión Europea
FAIR	Fácil de encontrar, accesible, interoperable y reutilizable
FRT	Tecnología de reconocimiento facial
FTC	Comisión Federal de Comercio
GAN	Redes generativas antagónicas
RGPD	Reglamento general de protección de datos
HUDERAF	Marco de Garantía de los Derechos Humanos, la Democracia y el Estado de derecho
ICCPR	Pacto Internacional de Derechos Civiles y Políticos
ICESCR	Pacto Internacional de Derechos Económicos, Sociales y Culturales
IFIP	Federación Internacional para el Tratamiento de la Información

ISO	Organización Internacional de Normalización
IoT	Internet de las Cosas
ITU	Unión Internacional de las Telecomunicaciones
LAPD	Departamento de Policía de Los Ángeles
LLM	Modelo de lenguaje de gran tamaño
MIT	Instituto Tecnológico de Massachusetts
ML	Aprendizaje automático
NDAS	Solución analítica de datos nacionales
ONG	Organización No Gubernamental
NIST	Institutos Nacionales de Normas y Tecnología
PNL	Procesamiento del lenguaje natural
OCDE	Organización para la Cooperación y el Desarrollo Económicos
SPC	Tribunal Popular Supremo
STF	Supremo Tribunal Federal
SUPACE	Portal de la Corte Suprema para la Asistencia en la Eficiencia de los Tribunales
TAR	Revisión asistida por la tecnología
UCL	Universidad Católica de Lovaina
DUDH	Declaración Universal de los Derechos Humanos
ONU	Naciones Unidas
UNESCO	Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura
EE. UU.	Estados Unidos

¿POR QUÉ ESTE KIT DE HERRAMIENTAS?

Este Kit de herramientas proporciona a los operadores judiciales el conocimiento y las herramientas necesarias para comprender los beneficios y riesgos de la Inteligencia Artificial ("IA") en su trabajo. El Kit de herramientas apoyará a los operadores judiciales en la reducción de los riesgos potenciales de la IA en relación con los derechos humanos al ofrecer orientación sobre las instancias, principios y regulaciones pertinentes del derecho internacional de los derechos humanos y la jurisprudencia emergente que sustenta el uso responsable de la IA.

El Kit de herramientas responde a la Recomendación de la UNESCO sobre la ética de la IA, adoptada por 193 países en 2021, que recomienda que "los Estados miembros deberían mejorar la capacidad del poder judicial para tomar decisiones relacionadas con los sistemas de IA según el Estado de derecho...".

¿Qué va a aprender?

Después de estudiar el kit de herramientas, los operadores judiciales podrán:

- Definir la IA y la Toma de decisiones algorítmicas (ADM) y entenderlas como sistemas sociotécnicos.
- Comprender los aspectos fundamentales relacionados con el sesgo algorítmico y la discriminación (como el sesgo de género, el sesgo racial y otras formas de sesgo que se cruzan) y explicar por qué son importantes en los entornos judiciales.
- Explicar el impacto de la IA en los siguientes derechos fundamentales: privacidad, libertad de expresión, acceso a la información, protección contra la discriminación, derecho al acceso a los tribunales, juicios y audiencias justos e imparciales y debido proceso legal.
- Examinar los casos judiciales relacionados con el uso de la IA, basándose en su conocimiento de las iniciativas regulatorias recientes y la jurisprudencia relacionada con el sesgo algorítmico, el uso inapropiado de algoritmos en la toma de decisiones.
- Aplicar herramientas como la Evaluación de impacto ético de la UNESCO para comprender el impacto ético de los sistemas de IA.

El Kit de herramientas se compone de cuatro módulos que completan un programa de capacitación sobre IA, derechos humanos y Estado de derecho para el poder judicial. El Kit de herramientas también proporciona los conocimientos necesarios no solo para los jueces, sino también para otros actores involucrados en el proceso contencioso, incluidos abogados y árbitros.

- ***Módulo 1: Introducción a la IA y al Estado de derecho***

  El módulo uno presenta al lector los conceptos principales relacionados con la el espaciado entre las palabras es menor que en el resto del documento? del desarrollo de la IA. El módulo define términos como IA, algoritmos, sistemas algorítmicos y describe sus características fundamentales y elementos básicos.

El módulo uno también analiza la importancia de los datos y la ciberseguridad en el contexto de la IA y proporciona una descripción general de los riesgos clave asociados con la IA, como las cajas negras.

- ***Módulo 2: Adopción de IA en el poder judicial***

  El módulo dos aborda la adopción de la IA en el poder judicial. Describe los usos de la IA en el poder judicial, como el descubrimiento electrónico y la revisión de documentos, el uso de IA generativa para ayudar con la redacción de documentos, el análisis predictivo y el soporte de ADM, las herramientas de evaluación de riesgos, la resolución de disputas, el reconocimiento y análisis de idiomas, el archivo digital y la gestión de casos. A continuación, el módulo destaca los casos de estudio sobre el despliegue de IA en el poder judicial en diferentes países y describe las oportunidades y los desafíos relacionados con estos ejemplos de uso.

- ***Módulo 3: Desafíos legales y éticos de la IA***

  El módulo tres presenta los principales desafíos legales y éticos relacionados con la IA en el poder judicial y resume las cuestiones legales relacionadas con la identificación biométrica y la tecnología de reconocimiento facial. El módulo tres analiza en detalle los desafíos relacionados con la IA y la ética basado en la Recomendación de la UNESCO 2021 relativa a la Ética de la inteligencia artificial.[1]

- ***Módulo 4: Derechos humanos e IA***

  El módulo cuatro presenta un análisis en profundidad de los derechos humanos afectados por la IA, como (i) el derecho al acceso a los tribunales, a un juicio justo y el debido proceso, (ii) a un recurso efectivo, (iii) los derechos a la protección contra la discriminación, (iv) a la libertad de expresión y el acceso a la información, y (v) el derecho a la privacidad y la protección de datos. El módulo también ofrece una visión general de los principales enfoques de gobernanza de la IA: basados en el riesgo y en los derechos humanos.

¿Quién se beneficiará de este Kit de herramientas?

El público objetivo principal del Kit de herramientas está formado por jueces, fiscales, procuradores, abogados públicos, universidades de derecho e instituciones de formación judicial.

¿Cómo utilizar este Kit de herramientas para la enseñanza?

El Kit de herramientas incluye actividades y recursos pertinentes a la IA, los derechos humanos y el Estado de derecho para los operadores judiciales. El Kit de herramientas se puede adaptar a las necesidades específicas de cada programa de formación judicial. El número de horas y la duración del programa de formación dependerán de la metodología elegida por el programa de formación judicial. El programa puede facilitarse como un programa de aprendizaje en línea, en el aula o híbrido, y puede ofrecerse como un curso intensivo o regular de un programa de pregrado, posgrado o educación continua, en función de la disponibilidad de formadores y/o la distribución geográfica de los alumnos matriculados para un curso específico, y el nivel de accesibilidad y conectividad.

Es mejor enseñar el programa como un esfuerzo organizado para transferir el conocimiento y desarrollar las habilidades y actitudes que fomenten acciones orientadas a la promoción

1 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000380455

y protección de los derechos humanos en relación con la IA. Por lo tanto, se recomiendan los siguientes elementos para cualquier formación basada en el Kit de herramientas:

- **Transferencia de conocimiento:** en el contexto de este Kit de herramientas, "conocimiento" se refiere a las normas de derechos humanos y los mecanismos de protección que son pertinentes para la IA para el grupo objetivo de alumnos. Por ejemplo, en el contexto de un curso donde el público objetivo son los jueces, el conocimiento podría referirse a los estándares de derechos humanos para las decisiones en casos que involucran el uso de IA.

- **Desarrollo de habilidades:** una comprensión básica de las normas de derechos humanos aplicables puede ser insuficiente para permitir que los alumnos conviertan estas normas en un comportamiento real. Las habilidades se afinan a través de la práctica, la aplicación y la reflexión, un proceso que puede iniciarse durante la capacitación a través de diversas actividades, pero que puede ser necesario continuar después del curso de capacitación, incluso a través de programas de seguimiento adecuadamente planificados. Por ejemplo, la capacidad de realizar una evaluación de riesgos de los sistemas de IA para determinar si deberían implementarse en primera medida, en lugar de asumir el despliegue y luego intentarla a posteriori para mitigar los daños.

- **Desarrollo de actitudes:** esto implica la adquisición y el refuerzo de actitudes positivas hacia los derechos humanos y el Estado de derecho, de modo que los alumnos tomen medidas para promover y salvaguardar los derechos humanos en su vida cotidiana y las responsabilidades profesionales al juzgar las violaciones de los derechos humanos que involucran los procesos de ADM y la IA.[2]

- El contenido de la capacitación se ha puesto a disposición en línea como un recurso abierto y puede actualizarse periódicamente mediante la creación de un repositorio en línea de presentaciones que los capacitadores pueden consultar y reutilizar bajo licencias abiertas de Creative Commons (Atribución 4.0 Internacional).[3]

2 Cualquier cambio que conduzca a un mejor respeto de los derechos humanos (cambios a nivel de los alumnos individuales, su organización/grupo y la comunidad/sociedad en general) que pueda atribuirse de manera plausible al esfuerzo de capacitación debe considerarse en una evaluación del impacto de la capacitación.

3 Véase: https://creativecommons.org/licenses/by/4.0/

Glosario

- **Datos agregados:** la agregación de datos implica recopilar una cantidad significativa de información de una base de datos y presentarla en un formato más manejable.
- **IA como una "caja negra":** el término "caja negra" se utiliza para denotar un sistema tecnológico que es inherentemente opaco, cuyo funcionamiento interno o lógica subyacente no se comprenden adecuadamente, o cuyos resultados y efectos no se pueden explicar.
- **Sesgo de IA:** el sesgo de IA es una diferencia sistemática en el tratamiento de ciertos objetos, personas o grupos (por ejemplo, estereotipos, prejuicios o favoritismo) en comparación con otros mediante algoritmos de IA.
- **Algoritmo:** un algoritmo se refiere a una serie de instrucciones para realizar cálculos u otras tareas, ya sea en matemáticas o en informática. En el caso de la IA, un algoritmo proporciona las instrucciones que permiten a un computador aprender a aprender del entorno y realizar un conjunto de tareas.
- **Toma de decisiones algorítmicas (ADM):** la toma de decisiones algorítmicas (ADM) se refiere al uso de "resultados producidos por algoritmos para tomar decisiones".
- **Etiquetado de datos:** el etiquetado de datos en el aprendizaje automático (ML) es el proceso de reconocer datos sin procesar (imágenes, archivos de texto, videos, etc.) y agregar una o más etiquetas relevantes y útiles para ofrecer contexto a fin de que un modelo de ML aprenda de él. Las etiquetas pueden mostrar si una fotografía contiene un pájaro o un automóvil, si las palabras se dijeron en una grabación de audio o si una radiografía muestra un tumor. Numerosos casos de aplicación necesitan etiquetado de datos, incluida la visión artificial, el procesamiento del lenguaje natural y el reconocimiento de voz.
- **Datificación:** el proceso de "datificación" se refiere a la proliferación de herramientas digitales utilizadas para integrar, analizar y mostrar patrones de datos.
- **Fideicomisos de datos:** una organización independiente que actúa como fideicomisario de los proveedores de datos y regula el uso adecuado de sus datos.
- **Deepfake:** un deepfake es cualquier forma de medio (video, audio u otro) que ha sido alterado o creado total o parcialmente desde cero.
- **Modelo de difusión:** los modelos de difusión son modelos generativos que están más avanzados que las redes generativas antagónicas (ver a continuación) en la síntesis de imágenes. Más recientemente, los modelos de difusión se utilizaron en DALL-E 2, el modelo de generación de imágenes de OpenAI y en Imagen de Google.

- **IA explicable (XAI):** la IA explicable (XAI) se define como sistemas, algoritmos y modelos con la capacidad de explicar su justificación para las decisiones, caracterizar las fortalezas y debilidades de su proceso de toma de decisiones y transmitir una comprensión de cómo se comportarán en el futuro.
- **Redes generativas antagónicas (GAN):** las GAN son un enfoque no supervisado de aprendizaje profundo que puede generar material hiperrealista. Las GAN se utilizan para técnicas de aprendizaje profundo no supervisadas, como la generación de imágenes realistas o conjuntos de datos de imágenes, la realización de traducciones de texto a imagen e imagen a texto, el envejecimiento de las caras y la creación de emojis.
- **IA generativa:** los algoritmos de aprendizaje automático (ML) se han diseñado para crear contenido nuevo, incluidos audio, código, imágenes, texto, simulaciones y videos.
- **Valor hash:** valores devueltos por una función hash, que se utiliza para convertir datos digitales de tamaño arbitrario en una cadena de salida con un número de caracteres de tamaño fijo.
- **Humano en el circuito (HITL):** HITL se refiere a un proceso en el cual un sistema de IA es monitoreado de cerca por un humano, quien es responsable de tomar todas las decisiones finales. Esto es especialmente importante en campos como la atención médica, donde la IA puede proporcionar un apoyo invaluable para hacer recomendaciones para el tratamiento del cáncer, la terapia de la sepsis, la planificación quirúrgica y más. Si bien las herramientas de IA pueden ayudar a los proveedores de atención médica a tomar decisiones informadas de manera rápida y precisa, la responsabilidad final de la atención al paciente siempre recae en el experto humano.
- **Aprendizaje automático (ML):** el ML es un conjunto de técnicas que permite a las máquinas aprender automáticamente utilizando patrones y deducciones en lugar de instrucciones directas de una persona. Las técnicas de ML con frecuencia instruyen a las máquinas para que alcancen un resultado al proporcionar numerosas instancias de resultados correctos. Sin embargo, también pueden especificar un conjunto de pautas y dejar que la máquina las descubra por sí misma en los datos.
- **Redes neuronales:** las redes neuronales son un tipo de técnica de aprendizaje automático que permite a los computadores aprender a realizar tareas mediante el análisis de ejemplos de entrenamiento. Por lo general, estos ejemplos están preetiquetados. Por ejemplo, un sistema de reconocimiento de objetos puede recibir miles de imágenes etiquetadas de objetos como automóviles, casas y tazas de café. A través del análisis, puede identificar patrones en las imágenes que correspondan a las etiquetas específicas. Una red neuronal está diseñada para parecerse a la estructura del cerebro humano, con miles o millones de nodos de procesamiento interconectados. Estos nodos suelen estar organizados en capas y los datos fluyen a través de ellos en una sola dirección, lo que los convierte en "alimentación prospectiva". Cada nodo recibe datos de los nodos de la capa inferior y envía datos a los nodos de la capa superior.

- **Procesamiento del lenguaje natural (PLN):** el PLN es una técnica de aprendizaje automático que analiza grandes cantidades de texto humano o datos del habla (transcritos o acústicos) en busca de propiedades específicas, como significado, contenido, intención, actitud y contexto.
- **Análisis predictivo:** el análisis predictivo es la categoría general de herramientas y modelos estadísticos, por ejemplo, sistemas de aprendizaje automático, que utilizan y analizan datos históricos para crear predicciones sobre el futuro para guiar la toma de decisiones. Estas predicciones pueden ser de bajo riesgo (por ejemplo, qué película recomendar), de riesgo medio (qué solicitud de préstamo proponer aceptar) o de alto riesgo (qué acusado tiene más posibilidades de involucrarse en un comportamiento en particular).
- **Discriminación por proxy:** la discriminación por proxy en los sistemas de IA ocurre cuando una característica aparentemente neutral se sustituye por una prohibida.
- **Aprendizaje automático supervisado:** el aprendizaje automático supervisado implica proporcionar a un sistema de aprendizaje automático un conjunto de datos que ya están etiquetados o clasificados, que el sistema puede usar para aprender a realizar una tarea en particular con precisión de acuerdo con las instrucciones dadas. El sistema de ML se carga con un conjunto de datos y el resultado esperado. En la fase de entrenamiento, el modelo ML ajusta sus variables para conectar las entradas con la salida correspondiente. Crear un algoritmo de aprendizaje supervisado exitoso requiere un equipo comprometido de especialistas para evaluar y analizar los resultados. Esto involucra a los científicos de datos que examinan a fondo los modelos producidos por el algoritmo para verificar su precisión frente a los datos de origen e identificar cualquier inexactitud causada por la IA.
- **Sandboxes regulatorios:** herramientas regulatorias que permiten a las empresas probar y experimentar con productos, servicios o negocios nuevos e innovadores bajo la supervisión de un regulador durante un período de tiempo limitado.


Technology
Innovation
SYSTEM
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
Technology
Innovation

Módulo 1

Introducción a la IA y al Estado de derecho

El módulo uno presenta la gobernanza algorítmica, los derechos humanos y el Estado de derecho. Analiza las definiciones de IA, algoritmos y sistemas algorítmicos, describiendo sus características principales y componentes básicos. El módulo subraya la importancia de los datos y la ciberseguridad en el contexto del despliegue de la IA en el poder judicial. Ofrece una visión general de los riesgos principales asociados con el despliegue de la IA en el poder judicial, como las cajas negras, y explica el principio humano en el circuito.

¿Qué va a aprender?

Después de completar este módulo, las personas participantes podrán:

- Comprender y explicar conceptos fundamentales relacionados con la IA, la gobernanza algorítmica y el Estado de derecho;
- Definir y explicar la IA, los algoritmos y los sistemas algorítmicos, describiendo sus características clave y sus componentes básicos;
- Comprender y reconocer los riesgos asociados con la IA, como las cajas negras y la ciberseguridad;
- Comprender la importancia del principio de humano en el circuito en el ciclo de vida de la IA;
- Comprender por qué los datos son importantes en el contexto de la IA.

1. Comprender la IA y sus componentes básicos

¿Qué son los sistemas de IA?

- Según la UNESCO, los sistemas de IA son sistemas que tienen la capacidad de procesar datos e información de una manera que se asemeja a un comportamiento inteligente, que generalmente incluye aspectos de razonamiento, aprendizaje, percepción, predicción, planificación o control.[4] En otras palabras, los sistemas de IA son tecnologías de procesamiento de información que integran modelos y algoritmos que producen una capacidad para aprender y realizar tareas cognitivas que conducen a resultados como la predicción y la toma de decisiones en entornos físicos y virtuales. Los sistemas de IA están diseñados para operar con diversos grados de autonomía mediante el modelado y la representación del conocimiento, y mediante la explotación de datos y el cálculo de correlaciones. Los sistemas de IA pueden incluir varios métodos, tales como (pero sin limitarse):

- aprendizaje automático, incluido el aprendizaje profundo y el aprendizaje por refuerzo;

- el razonamiento automático, incluida la planificación, la programación, la representación y el razonamiento del conocimiento, la búsqueda y la optimización.

Es importante tener en cuenta que tal definición tendría que cambiar con el tiempo, de acuerdo con los avances tecnológicos. Además, la IA a menudo se usa indistintamente con el término "aprendizaje automático" (ML), mientras que la IA es un campo mucho más amplio que se centra en muchas cosas más allá del ML, como la representación del conocimiento, la planificación y el razonamiento.[5]

Además de la descripción anterior, la Tabla 1 presenta una instantánea de cómo las diferentes organizaciones definen la IA de manera pragmática, de acuerdo con el conjunto de tareas o funciones que la tecnología puede realizar (OCDE, ISO), o de acuerdo con los ideales humanísticos que buscan incorporar en todo tipo de sistemas basados en datos para garantizar que contribuyan al mejoramiento de la sociedad (CE, UIT).

4 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000380455

5 OCDE (2019). Artificial Intelligence in Society, disponible en: https://www.oecd.org/publications/artificial-intelligence-in-society-eedfee77-en.htm; Leslie D., Burr C., Aitken M., Cowls J., Katell M., and Briggs, M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer, The Council of Europe, disponible en: https://ssrn.com/abstract=3817999 or http://dx.doi.org/10.2139/ssrn.3817999

Tabla 1. Definiciones de IA en organizaciones internacionales y multilaterales

Organización	Definición de IA
OCDE[6]	La IA es un sistema basado en máquinas que puede, para un conjunto dado de objetivos definidos por el ser humano, hacer predicciones, recomendaciones o tomar decisiones que influyan en entornos reales o virtuales. Cuando se aplica, la IA tiene siete casos de uso diferentes, también conocidos como patrones, que pueden coexistir en paralelo dentro del mismo sistema de IA.
ISO[7]	Sistema de ingeniería que genera resultados como contenidos, previsiones, recomendaciones o decisiones para un conjunto determinado de objetivos definidos por el ser humano.
CE[8]	La IA comprende sistemas que muestran un comportamiento inteligente analizando su entorno y emprendiendo acciones -con cierto grado de autonomía- para alcanzar objetivos específicos.
UIT[9]	La IA se refiere a la capacidad de un computador o de un sistema robótico habilitado por computador para procesar información y producir resultados de forma similar al proceso de pensamiento de los humanos en el aprendizaje, la toma de decisiones y la resolución de problemas. En cierto modo, el objetivo de los sistemas de IA es desarrollar sistemas capaces de abordar problemas complejos de forma similar a la lógica y el razonamiento humanos.

Sistemas de IA en nuestra vida diaria

La IA ya forma parte de nuestro día a día, seamos conscientes o no. Piense en su bandeja de entrada de correo electrónico: es posible que note que ciertos correos electrónicos terminan en su carpeta de correo no deseado, mientras que otros se clasifican como “sociales” o “promocionales”. ¿Cómo ocurre esto? ¿Sabía que Google ha implementado algoritmos de IA para categorizar y filtrar automáticamente los correos electrónicos? Estos algoritmos son programas entrenados para identificar elementos específicos dentro de un correo electrónico que indican que podría tratarse de correo no deseado. Cuando el algoritmo reconoce estos elementos, marca el correo electrónico como no deseado y lo traslada a esa carpeta. Si bien los algoritmos no son perfectos, se mejoran constantemente. Si encuentra un correo electrónico legítimo en su carpeta de correo no deseado, puede informar a Google que se ha etiquetado erróneamente como correo no deseado. Esta retroalimentación ayuda a mejorar la precisión del algoritmo.[10]

Otro ejemplo de un sistema de IA en nuestras interacciones diarias se presenta como un chatbot de servicio al cliente. Cuando escribe su pregunta, el chatbot utiliza un algoritmo para reconocer palabras clave y determinar el tipo de asistencia que necesita. A partir de la información existente y recién adquirida, el modelo de aprendizaje automático genera una respuesta adecuada. A medida que el chatbot interactúa con más clientes y recibe datos adicionales, mejora con el tiempo.[11]

6 OCDE (2019). Artificial intelligence and responsible business conduct, disponible en: https://mneguidelines.oecd.org/RBC-and-artificial-intelligence.pdf
7 ISO (2021). ISO/IEC DIS 22989, disponible en: www.iso.org/standard/74296.html
8 Comisión Europea (2018). Communication Artificial Intelligence for Europe, disponible en: https://digital-strategy.ec.europa.eu/en/library/communication-artificial-intelligence-europe
9 UIT (2018). Policy Considerations for AI Governance, disponible en: www.itu.int/en/ITU-T/studygroups/2017-2020/03/Documents/Shailendra%20Hajela_Presentation.pdf
10 Véase: https://dig.watch/technologies/artificial-intelligence
11 Bravo K. (2023). How Does AI actually work?, disponible en: https://blog.mozilla.org/en/internet-culture/how-does-ai-work/

Otros ejemplos de sistemas de IA cotidianos incluyen el motor de recomendaciones de Netflix para sugerir películas y programas de televisión basados en nuestras preferencias, o asistentes de voz como Siri y Alexa que nos ayudan con consultas simples.

Actividad: Preguntas de reflexión

1. ¿Qué viene a su mente cuando escuchas el término IA? Enumere sus connotaciones libremente y compárelas con un compañero. ¿Tuvieron alguna idea similar? ¿Cómo se exteriorizan posiblemente estas ideas en los discursos públicos dominantes sobre la IA?
2. Imagine el desarrollo tecnológico de las próximas tres décadas en los siguientes entornos (alternativamente, seleccione solo uno de ellos): hogar/familia, escuela, atención médica. ¿Qué procesos se han automatizado? ¿Cómo ha afectado la automatización al comportamiento, la interacción social y las experiencias de las personas?

Invite a las personas participantes de la capacitación a ver los siguientes videos.

Fuente: BBC, https://youtu.be/fvtrRGmv7aU

Fuente: OECD, https://youtu.be/6Y_ysDHn4uU

¿Qué es un algoritmo?

Un algoritmo se refiere a una serie de instrucciones para realizar cálculos u otras tareas, ya sea en matemáticas o en informática. En el caso de la IA, un algoritmo proporciona las instrucciones que permiten a un computador aprender a aprender del entorno y realizar un conjunto de tareas.[12]

Si bien un algoritmo general puede ser simple, los algoritmos de IA son más complejos.

Los algoritmos de IA están diseñados para aprender de los datos de entrenamiento, que pueden estar etiquetados o no etiquetados. El algoritmo utiliza esta información para mejorar sus capacidades y llevar a cabo sus tareas. Algunos algoritmos de IA son capaces de un aprendizaje continuo y pueden incorporar nuevas entradas de datos para refinar su proceso, mientras que otros requieren la intervención de un programador para optimizar su rendimiento.[13]

La toma de decisiones algorítmica (ADM) se refiere al uso de "resultados producidos por algoritmos para tomar decisiones".[14]

> Los algoritmos funcionan tomando un conjunto de entradas, como la edad de una persona, el lugar de residencia, el estado civil o los ingresos, y ejecutándolos a través de un conjunto de pasos que crean una salida o salidas, o decisiones, para esa persona o grupo, como la elegibilidad para un programa de asistencia financiera o la escuela pública asignada a un niño. Los algoritmos se utilizan en diversos sectores y propósitos, desde decisiones de atención médica, elegibilidad de beneficios públicos, planificación de infraestructura, asignación presupuestaria, entre otros sectores, con diversos grados de complejidad y entradas.

Olas de desarrollo de IA

Los sistemas de IA de la primera ola eran sistemas expertos o basados en reglas, donde un computador seguía una programación específica para generar resultados. Sin embargo, los sistemas de IA de la segunda ola, basados en el aprendizaje automático, aprenden de los datos de entrenamiento e infieren reglas para predecir resultados específicos. Los sistemas de IA de la tercera ola combinan las ventajas de las dos olas anteriores y tienen capacidades añadidas de poder responder al contexto en el que se utilizan y proporcionar a los usuarios explicaciones para su proceso de toma de decisiones.[15]

Las siguientes secciones explican y se centran en (i) sistemas expertos y programación tradicional y (ii) aprendizaje automático.

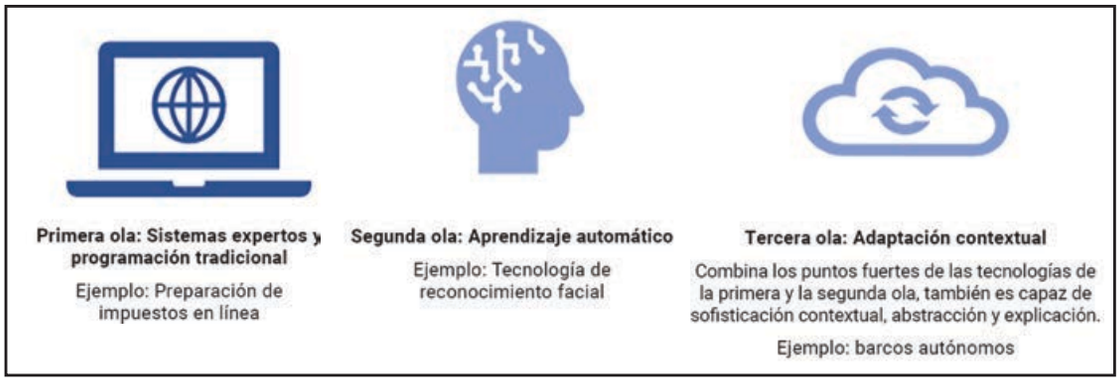
12 Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, disponible en: https://ssrn.com/abstract=4162985; Statement on Algorithmic Transparency Accountability, Association for Computing Machinery (2017), disponible en: https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf; also see: https://www.tableau.com/data-insights/ai/algorithms.

13 OCDE (2019). Artificial Intelligence in Society, disponible en: https://www.oecd.org/publications/artificial-intelligence-in-society-eedfee77-en.htm

14 Access Now (2018). Human rights in the age of artificial intelligence, disponible en: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

15 GAO (2021). Artificial Intelligence, An Accountability Framework for Federal Agencies and Other Entities, disponible en: https://www.gao.gov/products/gao-21-519sp.

Figura 1. Olas de desarrollo de IA



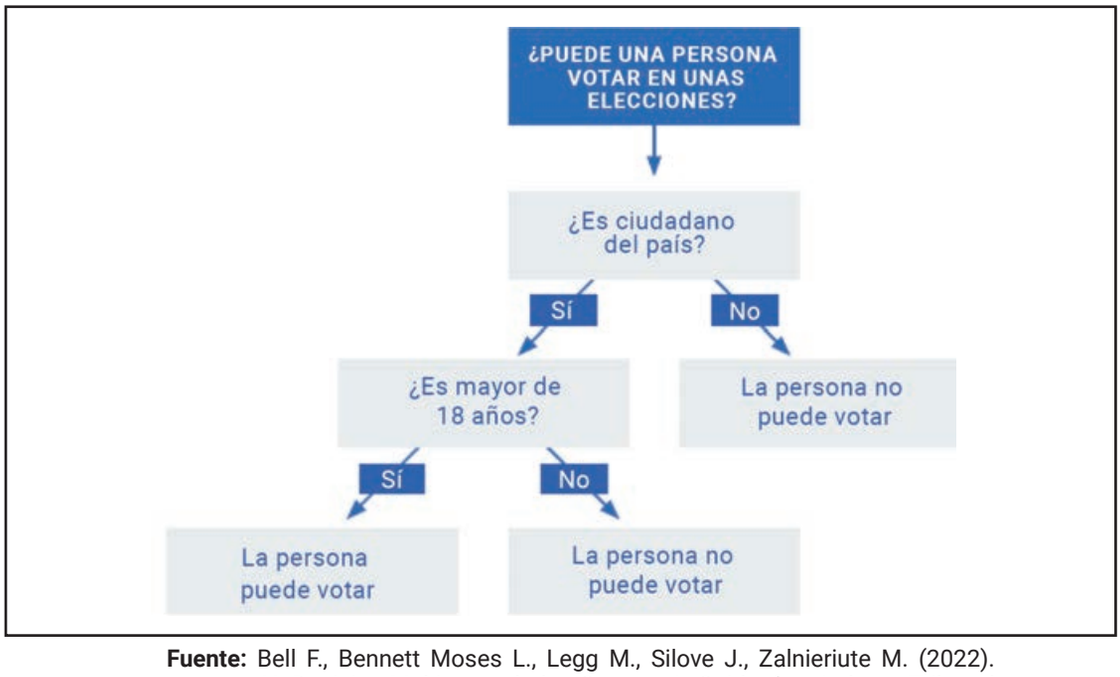
Fuente: Adaptado de GAO (2021). Artificial Intelligence, An Accountability Framework for Federal Agencies and Other Entities, disponible en: https://www.gao.gov/products/gao-21-519sp

Sistemas expertos y programación tradicional

Un "sistema experto" es un sistema de IA de "primera generación" que hace pronósticos, recomendaciones o conclusiones basadas en la entrada de datos. Implica una secuencia de etapas claramente programadas y las llamadas reglas "si..., entonces", que un computador puede aplicar para producir una salida. Estos sistemas suelen ser incapaces de tratar con información nueva o desafíos inesperados.

Las posibles opciones se denominan "nodos" en un árbol de decisiones, que es una representación visual de las reglas del sistema experto. La Figura 2 a continuación muestra un ejemplo de un árbol de decisiones que decide si una persona puede votar en una elección en un país donde los únicos requisitos previos para poder votar son que el individuo sea mayor de 18 años y ciudadano de un país en particular. Dado que cada rama solo tiene dos nodos, la Figura 2 es un ejemplo de un árbol de decisiones "binario".

Figura 2. Ejemplo de árbol de decisiones



Fuente: Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, disponible en: https://ssrn.com/abstract=4162985

Los sistemas expertos en IA de primera generación se utilizan ampliamente en los sistemas de planificación y optimización. Entre otros, los ejemplos incluyen software de procesamiento de impuestos, sistemas de servicio al cliente y soporte técnico, y sistemas de diagnóstico médico. Otro ejemplo es un método de alerta de fraude donde un experto especifica que si la información administrativa suministrada tiene más de cinco inexactitudes, el sistema debe emitir una alerta indicando que este expediente debe investigarse.

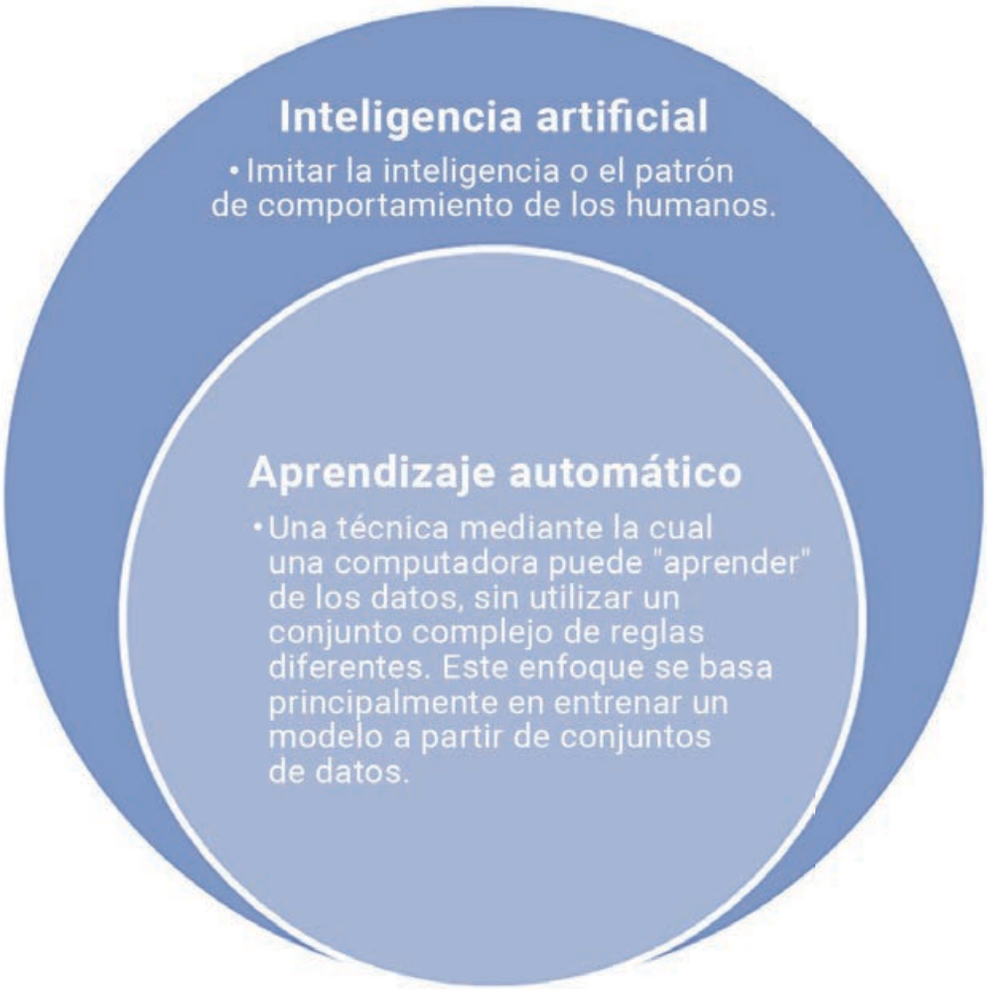
Inicialmente, era necesario dominar un lenguaje de programación para crear reglas en un lenguaje que una máquina pudiera entender. El concepto detrás de un "sistema experto" era que las reglas podrían desarrollarse por un experto en la materia (por ejemplo, un abogado) que no contara con habilidades de programación. Ahora hay una variedad de plataformas "sin código" disponibles que facilitan la "programación" de un computador para seguir un determinado procedimiento o llegar a conclusiones basadas en un conjunto de reglas. Ejemplos de tales plataformas incluyen Datalex de Austlii[16], Josef Legal[17], Checkbox[18], Neota Logic[19] y Realta Logic[20]. Estas plataformas permiten a los profesionales del derecho diseñar un conjunto de reglas utilizando, palabras, declaraciones, flechas, menús desplegables, u otros procesos similares, dependiendo de la plataforma que se utilice. Como resultado, incluso un abogado sin experiencia en programación puede codificar un árbol de decisiones como el que se muestra en la Figura 2.[21]

¿Qué es el aprendizaje automático?

Los sistemas de IA emplean cada vez más el aprendizaje automático (ML), que es un subconjunto de la IA. El aprendizaje automático es un conjunto de técnicas que permite a las máquinas aprender automáticamente utilizando patrones y deducciones en lugar de instrucciones directas de una persona.[22] Las técnicas de ML con frecuencia instruyen a las máquinas para que alcancen un resultado al proporcionar numerosas instancias de resultados correctos. Sin embargo, también pueden especificar un conjunto de pautas y dejar que la máquina las descubra por sí misma en los datos.[23]

16 Véase: https://austlii.community/wiki/DataLex
17 https://joseflegal.com/
18 https://www.checkbox.ai/
19 https://neota.com/
20 https://www.realtalogic.com/
21 Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, disponible en: https://ssrn.com/abstract=4162985
22 OCDE (2019). Artificial Intelligence in Society, disponible en: https://www.oecd.org/publications/artificial-intelligence-in-society-eedfee77-en.htm
23 Ibid. En ML se pueden encontrar numerosos métodos que han sido empleados por economistas, científicos e ingenieros durante años. Estos incluyen análisis de componentes principales, árboles de decisión, redes neuronales profundas y regresiones lineales y logísticas. Véase: https://www.oecd-ilibrary.org/sites/8b303b6f-en/index.html?itemId=/content/component/8b303b6f-en.

Figura 3. La relación entre IA y ML



Fuente: Autores

Existen muchas aplicaciones de ML. Algunas están diseñadas para un problema específico, como el reconocimiento de voz o de imágenes, mientras que otras se pueden utilizar para una gama más amplia de tareas.[24] El aprendizaje automático se ha integrado en los productos para abordar una variedad de problemas que son demasiado complicados para los sistemas de IA de "primera generación" o para la toma de decisiones humanas. El ML potencia los chatbots, el texto predictivo, las aplicaciones de traducción de idiomas, las recomendaciones de Netflix y la organización de las redes sociales. También permite vehículos autónomos y máquinas capaces de diagnosticar afecciones médicas mediante el análisis de imágenes.[25]

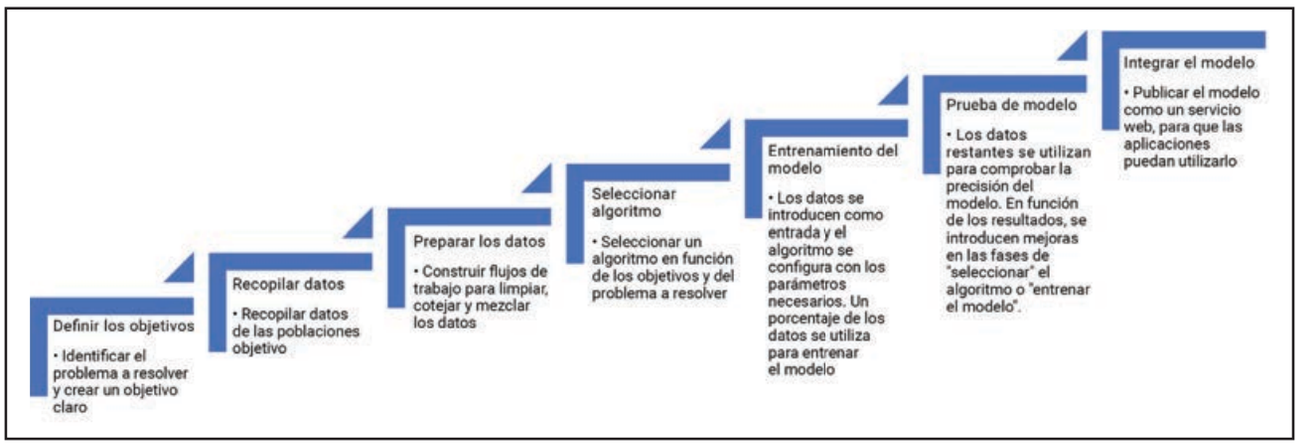
Los sistemas de ML "aprenden" a medida que analizan los datos. El ML es distinto del aprendizaje humano. Si bien ver pocas fotografías de un gato permitirá que un niño promedio comprenda el término "gato" y reconozca imágenes adicionales como gatos, los sistemas de ML requieren un conjunto de datos mucho más grande para realizar la misma tarea de

24 OCDE (2019). Artificial Intelligence in Society, disponible en: https://www.oecd.org/publications/artificial-intelligence-in-society-eedfee77-en.htm
25 Brown S. (2021). Machine learning, explained, disponible en: https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained.

categorización. El programa de ML se apoya en una base de datos que contiene imágenes de gatos y perros. Cada imagen está etiquetada con "gato" o "perro". Si al programa de ML se le muestran suficientes imágenes etiquetadas, el programa de ML comenzará a diferenciar las características de cada animal (entrenamiento o ajuste de ML). Una vez que el programa de ML aprenda, podrá adivinar a qué clase pertenece cada imagen. Se pueden realizar experimentos muy similares con texto.[26] Otro buen ejemplo de un programa de ML es el proceso de asignación de puntajes de crédito por parte de las instituciones financieras, donde los datos utilizados para entrenar el sistema de ML ya están clasificados como positivos o negativos según el historial crediticio del cliente.[27] Debemos recordar que la eficacia de los modelos de ML depende de la cantidad de datos de entrenamiento disponibles, la calidad del entrenamiento y los datos de entrada, y la cantidad de potencia de cálculo utilizada para construir el modelo.[28]

La Figura 4 a continuación ofrece una descripción simplificada de un proceso de ML, que consta de las siguientes fases: (i) definición de objetivos; (ii) recopilación de datos; (iii) preparación de datos; (iv) selección del algoritmo; (v) entrenamiento del modelo; (vi) prueba del modelo; y (vii) integración del modelo.

Figura 4. Descripción simplista del proceso de ML



Fuente: Autores

26 Medvedeva M., Vols M., Wieling M. (2020). Using machine learning to predict decisions of the European Court of Human Rights, Artif Intell Law, 28, 237–266, disponible en: https://link.springer.com/article/10.1007/s10506-019-09255-y.

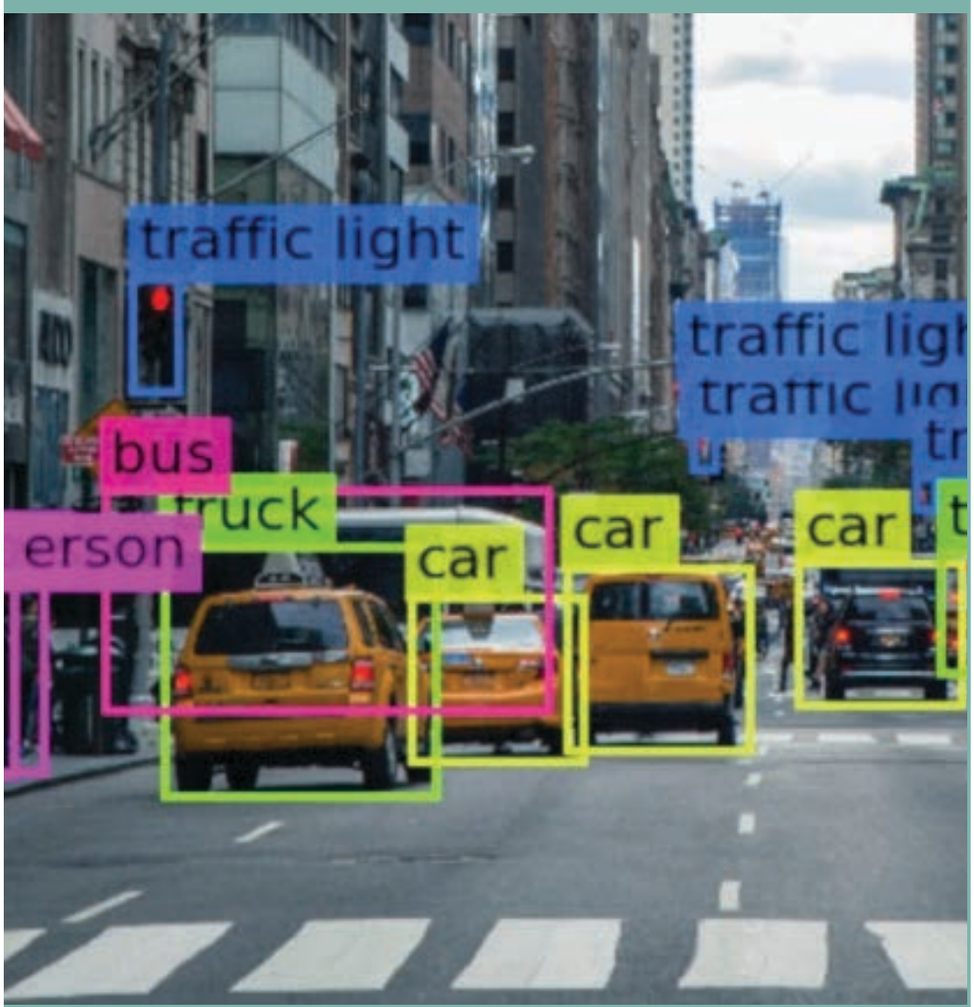
27 The Royal Society (2012). Machine Learning: The Power and Promise of Computers that Learn by Example, disponible en: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf 16; Allens Linklaters (2018). AI Toolkit: Ethical, Safe, legal; Practical Guidance for AI Projects, disponible en: https://lpscdn.linklaters.com/-/media/files/insights/thought-leadership/ai-toolkit/ethical-safe-lawful-toolkit-for-artificial-intelligence-projects-nov2018.ashx?rev=b82597fb-d88a-457d-a41a-a24ec1fc7253&extension=pdf) 9; https://humanrights.gov.au/our-work/rights-and-freedoms/publications/human-rights-and-technology-final-report-2021.

28 Stankovich M., Behrens E., Burchell J. (2023). Toward Meaningful Transparency and Accountability of AI Algorithms in Public Service Delivery, disponible en: https://www.dai.com/uploads/ai-in-public-service.pdf

Etiquetado de datos en ML

El etiquetado de datos en el Aprendizaje automático (ML) es el proceso de reconocer datos sin procesar (imágenes, archivos de texto, videos, etc.) y agregar una o más etiquetas relevantes y útiles para ofrecer contexto para que un modelo de ML aprenda de él. Las etiquetas pueden mostrar si una fotografía contiene un pájaro o un automóvil, si las palabras se dijeron en una grabación de audio o si una radiografía muestra un tumor. Numerosos casos de aplicación necesitan etiquetado de datos, incluida la visión por computador, el procesamiento del lenguaje natural y el reconocimiento de voz[29].

Figura 5. Etiquetado de datos en ML



Ejemplo de etiquetado de datos.

Fuente: Energy (2021). The One, Two, Threes of Data Labeling for Computer Vision, disponible en: https://medium.com/unpackai/the-one-two-threes-of-data-labeling-for-computer-vision-4c0b022cef4

29 Véase: https://aws.amazon.com/sagemaker/data-labeling/what-is-data-labeling/

El proceso de descubrimiento en litigios puede servir como un gran ejemplo de mostrar la complejidad de usar ML en el poder judicial. Véase la Figura 6.

Figura 6. Descubrimiento en litigios: tres posibles niveles de automatización

I nivel - Sin automatización alguna. En esta fase, un asistente jurídico examina los documentos jurídicos siguiendo una serie de parámetros predeterminados.

II nivel - Automatización sin ML. Un sistema informático utiliza criterios fijos, como intervalo de fechas, listas de frases, ubicación de archivos, para llevar a cabo la búsqueda de documentos.

III nivel - ML. Un asistente jurídico etiqueta los documentos que formarán parte del descubrimiento (estos son los datos de entrenamiento). A continuación, se puede utilizar un sistema de ML para inferir los criterios de búsqueda basándose en patrones de los datos de formación etiquetados por el ser humano, en lugar de utilizar únicamente a un ser humano para identificar qué características son necesarias para la descubribilidad. El modelo ML entrenado clasificará los documentos restantes en aquellos que son y no son susceptibles de ser descubiertos utilizando estos patrones.

Fuente: Autores

Actividad: las personas participantes en la capacitación analizan el siguiente escenario hipotético sobre el uso de pruebas generadas por IA en los procedimientos judiciales. ¿Qué haría si estuviera en una situación similar? ¿Qué cuestiones legales fundamentales tendrá en cuenta?

En un futuro no muy lejano, la evidencia generada por la IA juega un papel fundamental en un caso judicial de alto perfil. Así es como se desarrolla:

Antecedentes del caso: se acusa a una destacada empresa de tecnología de utilizar algoritmos sesgados en su proceso de contratación, lo que resulta en discriminación contra ciertos grupos demográficos. El caso ha atraído una gran atención pública y está siendo vigilado de cerca por sus posibles consecuencias en la ética de la IA y la responsabilidad corporativa.

Evidencia generada por IA:

1. **Informe de auditoría algorítmica:** los demandantes han empleado un equipo de especialistas en ética de IA y científicos de datos para llevar a cabo una auditoría exhaustiva de los algoritmos de contratación de la empresa. Presentan un informe detallado generado por los sistemas de IA que destaca los casos de sesgo y discriminación en el proceso de toma de decisiones del algoritmo.
2. **Simulación generada por IA:** para demostrar el comportamiento del algoritmo, los demandantes introducen una simulación generada por IA que imita el proceso de contratación de la empresa. Esta simulación utiliza datos históricos para mostrar cómo el algoritmo tiende a favorecer a ciertos grupos demográficos sobre otros.
3. **Testimonio de expertos generado por IA:** la defensa llama a un experto en ética de IA que utiliza IA de procesamiento de lenguaje natural para analizar las comunicaciones y documentos internos de la empresa. La IA identifica casos en los que los empleados expresaron su preocupación por el sesgo algorítmico, lo que podría sugerir que la empresa estaba al tanto del problema.

Consecuencias legales: la introducción de pruebas generadas por la IA presenta varios desafíos y consideraciones legales:

1. **Admisibilidad:** el tribunal debe determinar la admisibilidad de las pruebas generadas por IA, evaluando su fiabilidad y relevancia para el caso.
2. **Testimonio de expertos:** el tribunal se enfrenta a la cuestión de si la IA puede considerarse un "testigo experto" y cómo debe tratarse su testimonio.
3. **Consecuencias éticas:** el caso plantea cuestiones éticas sobre la responsabilidad de las empresas al implementar sistemas de IA y las posibles consecuencias del sesgo algorítmico.
4. **Impacto en el precedente:** el resultado de este caso podría sentar un precedente sobre cómo se tratan las pruebas generadas por IA en futuros procedimientos legales, influyendo en el panorama legal con respecto a la ética de la IA.
5. **Supervisión humana:** a pesar de la evidencia generada por IA, el juicio humano sigue siendo crucial para interpretar la evidencia, garantizar la imparcialidad y tomar decisiones legales.

Este escenario hipotético subraya la evolución del papel de la IA en los procedimientos legales, así como la necesidad de marcos legales sólidos para abordar las complejidades y las preocupaciones éticas asociadas con la evidencia generada por IA en los tribunales.

2. ¿Por qué son importantes los datos en el contexto de la IA?

Los algoritmos de IA requieren acceso a máquinas de datos que no pueden "aprender" a menos que tengan grandes conjuntos de datos a partir de los cuales establecer patrones. La disponibilidad de datos es un requisito necesario para el desarrollo de la IA que le permita realizar ciertas tareas previamente realizadas manualmente por humanos.

El proceso de "datificación" se refiere a la proliferación de herramientas digitales utilizadas para la integración, análisis y visualización de patrones de datos. La datificación indica que numerosos aspectos de la vida social asumen la forma de huellas digitales. Las amistades se convierten en "me gusta" en Facebook, los movimientos en toda la ciudad dejan vastas huellas digitales en dispositivos con GPS y las búsquedas de información revelan lo que las personas y las comunidades valoran o desean.[30]

Una vez que los dispositivos conectados a Internet comienzan a comunicarse entre sí, la mayoría de los usuarios envían una cantidad extraordinaria de datos nuevos sin saberlo y prácticamente sin que lo noten. Por ejemplo, hay metadatos (datos sobre datos), como la información de enrutamiento contenida dentro de los encabezados de correos electrónicos o mensajes de texto, o la información de geolocalización oculta dentro de una fotografía digital. Los metadatos, como información estructurada, pueden compararse y evaluarse más fácilmente mediante algoritmos y, por lo tanto, con frecuencia pueden proporcionar información inusualmente exacta sobre los intereses, movimientos y relaciones de los individuos.

Las plataformas digitales tienen acceso a mucha información sobre lo que las personas están haciendo en línea. Estos flujos masivos de rastros digitales, llamados big data, se pueden usar junto con técnicas de clasificación automatizadas, como algoritmos e IA, para revelar patrones importantes y conducir a información analítica sobre clientes, enfermedades y actividades delictivas. Muchas plataformas y empresas digitales buscan captar clientes desde el principio convirtiéndose en el lugar donde las personas compran libros o transmiten películas, por ejemplo. También quieren construir ecosistemas cerrados, como Netflix o Amazon, donde puedan controlar y extraer valor de los datos.[31]

La calidad de los datos afecta el resultado de la IA en términos de sesgo [para el sesgo de IA, consulte el Módulo 3]. Idealmente, los datos deben estar libres de sesgos, la propiedad de los datos debe estar claramente establecida y los algoritmos deben ser lo suficientemente transparentes como para indicar la responsabilidad de las partes interesadas. Las obligaciones de todas las partes interesadas en el ciclo de vida de la IA deben definirse para prevenir daños y reparar o compensar los daños causados por los sistemas de IA.

Al decidir casos que involucran el despliegue de IA y su impacto en los derechos humanos, los operadores judiciales deben considerar las siguientes preguntas relacionadas con los datos y conjuntos de datos que se incorporan a los sistemas de IA (ver Tabla 2).

30 Matteson A. (2018). The Concept of Datafication; Definition & Examples, disponible en: https://www.datasciencecentral.com/the-concept-of-datafication-definition-amp-examples/

31 Flyverbom M., Deibert R., Matten, D. (2019). The Governance of Digital Technology, Big Data, and the Internet: New Roles and Responsibilities for Business. Business & Society, 58(1), 3–19, disponible en: https://doi.org/10.1177/0007650317727540

Tabla 2. Preguntas relacionadas con datos y conjuntos de datos que se incorporan a los sistemas de IA.

Preguntas	Aspectos a tener en cuenta
Acceso a los datos y disponibilidad	La ausencia de sistemas necesarios que generen y mantengan datos sólidos, precisos y relevantes ha hecho que el desarrollo de aplicaciones de IA sea un desafío en algunos contextos.
Precisión de datos	El acceso a datos precisos es crucial para implementar con éxito la IA y los recursos digitales. Una buena práctica para salvaguardar la precisión de los datos es la práctica del llamado «degüelle algorítmico», que requiere que los desarrolladores de sistemas de IA eliminen cualquier dato que se haya obtenido ilegalmente y se haya utilizado para entrenar a los sistemas de IA.[32]
Calidad de los datos	Uno de los impedimentos clave para el despliegue efectivo de la IA en el poder judicial es el acceso a datos FAIR (localizables, accesibles, interoperables y reutilizables). Este problema se agrava en ciertos contextos porque los datos no siempre se digitalizan y no son de fácil acceso. Preguntas clave a realizar a este respecto: ¿cuál es la calidad de los datos en los que está entrenado el sistema de IA? ¿Existe el riesgo de sesgo de datos y amplificación de información incorrecta utilizando IA? El problema es que los datos que alimentan los sistemas de IA pueden ser inexactos, incompletos o contener errores o material sin importancia. Los datos pueden estar impregnados de sesgos. Muchas veces, las máquinas ya están recopilando datos sesgados que provienen de una realidad errática y sesgada. Por ejemplo, los ensayos clínicos a menudo excluyen a las mujeres y las personas de color, lo que lleva a una representación inadecuada de los datos. Esto podría tener graves consecuencias si se utilizan algoritmos entrenados con dichos datos para analizar imágenes de la piel o priorizar la atención a los pacientes. Como resultado, es crucial garantizar que los algoritmos de IA estén capacitados utilizando datos representativos para evitar tales sesgos y garantizar resultados equitativos para todos.[33] Ejemplo: la mayoría de los sistemas de IA utilizados en la justicia penal son modelos estadísticos, basados en datos de aplicación de la ley o antecedentes penales que representan sesgos estructurales y desigualdades sociales. Estos datos son un registro de los delitos, ubicaciones y grupos vigilados, y no son un registro necesario de la ocurrencia real del delito. Estos datos utilizados en los sistemas de IA pueden reforzar y volver a introducir patrones de discriminación en los sistemas judiciales o policiales.[34] Los reguladores de los modelos de IA deben asegurarse de que los datos utilizados se adhieran a los principios FAIR y se recopilen de manera ética antes de certificar el modelo como apto para el mercado. Esto podría complementarse con una evaluación de la calidad organizativa en los puntos de control previos a la comercialización. Estas condiciones pueden indicar a la industria que la integridad de los datos y la recopilación ética son de suma importancia para colocar soluciones de IA en el mercado y conducir a cambios estructurales positivos en el funcionamiento de las empresas.

32 La FTC utilizó esta práctica para obligar a Everalbum, creadores de la ya desaparecida aplicación Ever, a eliminar los sistemas de reconocimiento facial que se desarrollaron utilizando el contenido obtenido de los usuarios de la aplicación. Véase también: Kay K. (2021). Why the FTC is forcing tech firms to kill their algorithms along with ill-gotten data, disponible en: https://digiday.com/media/why-the-ftc-is-forcing-tech-firms-to-kill-their-algorithms-along-with-ill-gotten-data/

33 Siwicki B. (2021). How does bias affect healthcare AI, and what can be done about it?, disponible en: https://www.healthcareitnews.com/news/how-does-bias-affect-healthcare-ai-and-what-can-be-done-about-it

34 Fair Trials (2021). Regulating Artificial Intelligence for Use in Criminal Justice Systems in the EU Policy Paper, disponible en: https://www.fairtrials.org/sites/default/files/Regulating%20Artificial%20Intelligence%20for%20Use%20in%20Criminal%20Justice%20Systems%20-%20Fair%20Trials.pdf

Preguntas	Aspectos a tener en cuenta
Representatividad de los datos	Un conjunto de datos es representativo si refleja o mide con precisión la población o el fenómeno que debe registrar, en relación con su aplicación prevista.[35] Ejemplo: la dependencia excesiva de técnicas de recopilación de datos «automatizadas» puede dejar fuera a grupos extremadamente vulnerables y erosionar la confianza en la toma de decisiones automatizada. Las personas sin acceso digital (es decir, aquellas sin conectividad o dispositivos) o que carecen de habilidades digitales no serán consideradas en los análisis de la población y sus requisitos. Las brechas digitales en muchos países del hemisferio sur han llevado a la «invisibilidad de los datos», que probablemente afectará a grupos históricamente marginados como mujeres, castas, comunidades tribales, minorías religiosas y lingüísticas y mano de obra migrante. La utilidad y validez de los algoritmos de IA desarrollados a partir de datos fácilmente disponibles pueden verse limitadas por sesgos perpetuados por la invisibilidad de los datos. Esto subraya los requisitos de transparencia y responsabilidad algorítmica.
Propiedad de los datos	Un tema fundamental en el desarrollo y la implementación de la IA es la propiedad de los datos, es decir, quién posee, gestiona y recopila los datos que entran en el sistema de IA. Las cuestiones importantes a tener en cuenta en este sentido son definir los objetivos (por qué necesitamos el sistema de IA) y determinar qué datos de capacitación adquirir y cómo categorizar los datos. Por lo tanto, se requieren constantemente juicios humanos al compilar conjuntos de datos y desarrollar algoritmos para la predicción.
Almacenamiento de datos/ minimización de datos	El almacenamiento a largo plazo de datos personales conlleva riesgos, ya que los datos son susceptibles de explotación de formas que no se previeron en el momento de la recopilación de los datos. Los datos pueden volverse obsoletos, irrelevantes o contener malas interpretaciones históricas a lo largo del tiempo, lo que podría conducir a resultados sesgados o incorrectos del procesamiento de datos en el futuro.[36]
Protección de datos y privacidad	Las leyes de protección de datos adecuadas abordan cuestiones como la privacidad de los datos (un derecho humano básico), la gestión y el intercambio de datos y los mecanismos innovadores para el gobierno de los datos, como las sandboxes de datos y los fideicomisos de datos. Las políticas y regulaciones de datos actuales entre países y regiones están muy fragmentadas, con enfoques regulatorios globales, regionales y nacionales divergentes. Muchos países y regiones han tomado medidas para actualizar las normas sobre el uso de datos personales. El Reglamento General de Protección de Datos[37] (RGPD) de la UE impone una larga lista de requisitos a las empresas que tratan datos personales. Las infracciones resultan en multas que podrían totalizar hasta el 4 % de la facturación anual global. El RGPD permite un mejor control de los datos personales, lo que permite la protección individual del anonimato, el seudónimo y el derecho al olvido. La portabilidad de los datos otorga a las personas el derecho a solicitar que sus datos se transfieran a otro controlador y que los controladores de datos utilicen formatos comunes. Más del 30 % de los países, principalmente los países en desarrollo, no tienen legislación sobre gobernanza de datos, y pocos han desarrollado una ley integral de protección de datos[38]. Otros marcos regionales para establecer normas sobre la privacidad de los datos personales incluyen el Marco de Privacidad de APEC (2015); y las Directrices de Privacidad de la OCDE (2013) y el Convenio 108+ del Consejo de Europa[39], que ha actualizado las directrices sobre protección de datos. Vale la pena que en los países que no cuentan con un sistema de protección de datos, los tribunales tengan que establecer pautas para el uso de los datos, lo que estaría en consonancia con los derechos legales.

35 AAAS. AAAS, Artificial Intelligence and the Courts: Materials for Judges, disponible en: https://www.aaas.org/ai2/projects/law/judicialpapers

36 Consejo de Derechos Humanos de las Naciones Unidas (2021). The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, disponible en: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

37 Guía completa sobre el cumplimiento del RGPD, disponible en: https://gdpr.eu/

38 UNCTAD, Data Protection and Privacy Legislation Worldwide, disponible en: https://unctad.org/page/data-protection-and-privacy-legislation-worldwide

39 Council of Europe, Modernisation of Convention 108, disponible en: https://www.coe.int/en/web/data-protection/convention108/modernised

Preguntas	Aspectos a tener en cuenta
Infraestructura de datos	El progreso actual en IA y big data se ve impulsado por mejores conexiones digitales, cantidades crecientes de datos, algoritmos sofisticados y una mayor potencia de procesamiento. La IA y big data pueden mejorar enormemente la vida en los países en desarrollo y ayudar a alcanzar los Objetivos de desarrollo sostenible de las Naciones Unidas. Los responsables políticos deben apuntar a permitir, incentivar y/o acelerar la inversión en la construcción de una infraestructura de datos adecuada y asequible. Se necesita invertir en software, hardware y conectividad de banda ancha para un acceso y uso generalizados de los datos. Esto es fundamental para llegar al segmento desatendido. Incentivar la creación de datos FAIR e infraestructura de datos FAIR es fundamental.[40]
Preguntas adicionales para hacer	• ¿El sistema de IA se sometió a auditorías de transparencia algorítmica o evaluaciones de impacto en la privacidad? • ¿Se utilizaron técnicas de mejora de la privacidad para preservar la privacidad de los datos? • ¿Cuál es el estado de la información y la ciberseguridad para la privacidad de los datos?

3. Sistemas de IA como "cajas negras"

El término "caja negra" se utiliza para denotar un sistema tecnológico que es inherentemente opaco, cuyo funcionamiento interno o lógica subyacente no se comprenden adecuadamente, o cuyos resultados y efectos no se pueden explicar.[41] Muchos sistemas de IA se consideran "cajas negras", es decir, sistemas altamente complejos cuyos procesos de toma de decisiones y razonamiento no son fáciles de entender por los usuarios, e incluso por sus desarrolladores. Esto puede hacer que sea extremadamente difícil detectar salidas defectuosas, particularmente en sistemas de IA que descubren patrones en los datos subyacentes de manera no supervisada.

Los sistemas de IA analizan los datos de entrenamiento para identificar patrones complejos y luego aprenden estos patrones para clasificar los nuevos datos que pueden recibir. Sin embargo, muchos sistemas de IA no explican cómo los datos podrían estar interrelacionados y cómo llegan a una determinada decisión o predicen un determinado resultado. Estos sistemas pueden ser demasiado complejos para la comprensión humana, incluso para aquellos que los programan y entrenan.[42] Evolucionan y aprenden continuamente y tienen un comportamiento impredecible. Es posible que puedan deducir hechos y correlaciones a partir de variables indirectas, como el historial de compras o la geografía.

40 FAIR Principles, disponible en: https://www.go-fair.org/fair-principles/
41 AAAS, Artificial Intelligence and the Courts: Materials for Judges, disponible en: https://www.aaas.org/ai2/projects/law/judicialpapers
42 OCDE, AI in Society, disponible en: https://www.oecd-ilibrary.org/sites/969ff07f-en/index.html?itemId=/content/component/969ff07f-en

En profundidad: discriminación por proxy en sistemas de IA

La discriminación por proxy en los sistemas de IA tiene lugar cuando una característica aparentemente neutral se sustituye por una prohibida.[43]

Por ejemplo, las instituciones financieras a menudo usan códigos postales y límites de vecindario (geografía), estos datos pueden capturar la raza de los solicitantes de préstamos, ya que algunos códigos postales pueden estar asociados con grupos sociales de bajos ingresos, minorías étnicas o raciales. Del mismo modo, un sistema de IA creado por una compañía de seguros puede aumentar las primas para los solicitantes que pueden ser miembros de un grupo de Facebook dedicado a mejorar la disponibilidad de pruebas genéticas de predicción del cáncer. En estas condiciones, es probable que la aseguradora esté incurriendo en discriminación genética indirecta mediante el uso de proxies, como la demanda de cierto tipo de pruebas genéticas y la membresía a un grupo específico de Facebook, para deducir el vínculo entre estos proxies y la historia genética (una práctica controvertida) y cobrar primas de seguro más altas a dichas personas.[44] Otro ejemplo serían los proxies relacionados con la edad, "veinte años de experiencia profesional" indica que la persona debe tener al menos cuarenta y tantos años.

Los derechos a la privacidad y la no discriminación en los sistemas automatizados de toma de decisiones exigen la minimización, limitación o prohibición de ciertos usos de los datos, o la eliminación de datos (consulte la Tabla 2 anterior). Sin embargo, un sistema de IA puede hacer una predicción basada en datos proxy que se parezca mucho a las categorías restringidas de datos. Además, la única forma de descubrir estos proxies es adquirir información sensible o privada como la raza. Si se adquieren dichos datos, es crucial garantizar que se utilicen exclusivamente para fines adecuados y legítimos.[45] Por ejemplo, aunque los creadores de algoritmos pueden haber hecho un esfuerzo consciente para evitar el sesgo racial al excluir la raza como parámetro, el algoritmo producirá resultados que están sesgados racialmente si incluyen proxies típicos de raza, como ingresos, educación o código postal.

La opacidad de los algoritmos de IA y la dificultad para determinar la responsabilidad por las decisiones producidas por los sistemas de IA significan que pueden ocurrir daños a los derechos humanos, y no se establece ninguna responsabilidad por estos daños. Sin la incorporación de salvaguardias éticas y de derechos humanos en el diseño y la implementación de la IA, los riesgos relacionados con la IA se intensificarán. Esto tendrá un impacto en la profundización de las desigualdades existentes incrustadas en los conjuntos de datos utilizados para entrenar algoritmos. Por ejemplo, estas desigualdades podrían derivarse del sesgo de los desarrolladores. Esto afectará de manera severa y desproporcionada a los grupos desfavorecidos, desatendidos y marginados, y a aquellos que están sujetos a formas interseccionales de discriminación.

43 Downs J., Auchterlonie S. (2022). Proxy Problems—Solving for Discrimination in Algorithms, disponible en: https://www.bhfs.com/insights/alerts-articles/2022/proxy-problems-solving-for-discrimination-in-algorithms

44 Iowa Law Review (2020). Proxy Discrimination in the Age of Artificial Intelligence and Big Data, disponible en: https://ilr.law.uiowa.edu/print/volume-105-issue-3/proxy-discrimination-in-the-age-of-artificial-intelligence-and-big-data

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

Otro problema es el uso indebido de las salvaguardias de la propiedad intelectual. Las herramientas algorítmicas a menudo caben bajo el manto de software patentado y reclamos de secretos comerciales para proteger la tecnología detrás de los algoritmos del escrutinio externo (ver People vs. Chubbs analizado en el Módulo 4 a continuación). Esta práctica podría impedir cualquier esfuerzo de defensa para desafiar la fiabilidad de la ciencia subyacente a la herramienta de IA. Cuando los sistemas de IA se utilizan en operaciones en nombre de las partes interesadas del sistema de justicia, existe una necesidad acentuada de rendición de cuentas, transparencia y explicación. Las salvaguardias de propiedad intelectual de los datos y el sistema algorítmico pueden impedir dicha transparencia y responsabilidad. Las partes interesadas en la gobernanza de la IA deberán encontrar un equilibrio entre la transparencia como parte de la ética de la IA y la necesidad legítima de proteger los secretos comerciales cuando las empresas privadas desarrollen herramientas de IA.

Actividad: secretos comerciales, algoritmos y derechos fundamentales: el caso de estudio del algoritmo del Sistema de Evaluación del Valor Añadido Educativo (EVAAS)

Los secretos comerciales que protegen los algoritmos afectan a los derechos fundamentales. Lea el estudio de caso a continuación y analice cómo se juzgaría un caso similar en su país. ¿Cómo se decidiría este caso en virtud de la legislación nacional?

Entre 2011 y 2015, el desempeño laboral de los maestros de Houston se evaluó utilizando un algoritmo "basado en datos": EVAAS. El programa permitió a la junta de educación automatizar las opciones sobre si los maestros recibían bonificaciones, eran penalizados por su bajo rendimiento o incluso despedidos. Los códigos fuente son secretos comerciales propiedad de SAS, un proveedor externo. Como tal, los maestros no pudieron impugnar las decisiones ni recibir una explicación de cómo el EVAAS llegó a sus decisiones.

Se produjo un largo litigio civil, y en 2017, un juez federal de EE. UU. concluyó que los derechos constitucionales de los demandantes fueron violados por el despliegue del algoritmo secreto para evaluar el desempeño de los empleados sin una explicación adecuada. El juez tuvo que encontrar un equilibrio entre el derecho comprensible del vendedor privado a preservar sus secretos comerciales y el derecho constitucional de los maestros al debido proceso, que protege a los ciudadanos estadounidenses de las privaciones de la vida, la libertad o la propiedad que son fundamentalmente injustas o erróneas.

La decisión judicial declaró que los maestros y la Federación de Maestros de Houston deben poder verificar e impugnar de forma independiente los resultados de la evaluación producidos por el algoritmo. Sin embargo, SAS se negó a revelar cómo funciona su algoritmo EVAAS internamente. Como resultado, el sistema escolar de Houston ya no utiliza el algoritmo EVAAS.

Fuente: Hung K-H., Liddicoat J. (2018). The future of workers' rights in the AI age, disponible en: https://policyoptions.irpp.org/magazines/december-2018/future-workers-rights-ai-age/

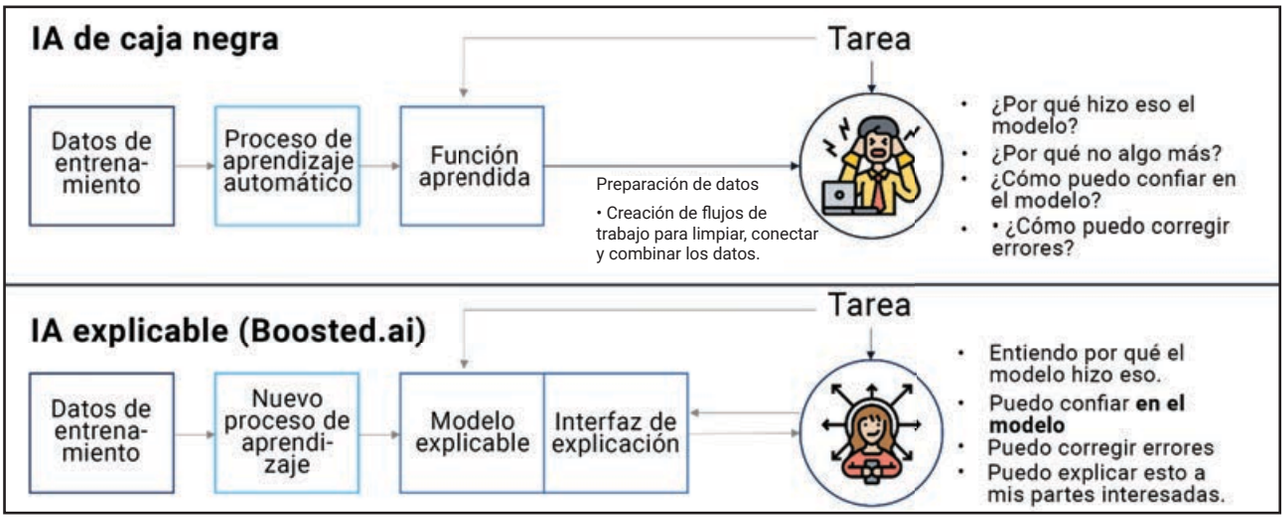
IA explicable (XAI)

El debate sobre los aspectos de la caja negra de los sistemas de IA está en continua evolución. Los avances en la investigación de la IA han llevado al desarrollo de modelos de IA que no son cajas negras.

La IA explicable (XAI) se define como sistemas, algoritmos y modelos con la capacidad de explicar su justificación para las decisiones, caracterizar las fortalezas y debilidades de su proceso de toma de decisiones y transmitir una comprensión de cómo se comportarán en el futuro.

Los investigadores en XAI se concentran en crear modelos de IA que puedan comprenderse por las personas, así como en producir explicaciones de los resultados de ML que sean utilizables. Esta audiencia debe tener la oportunidad de analizar el modelo generado y discernir su significado, es decir, comprender la estructura del sistema.

Figura 7. La IA de caja negra frente a la IA explicable



Fuente: https://boosted.ai/

Por ejemplo, Angelino et al (2018) desarrollaron un modelo de ML interpretable para pronosticar el nuevo arresto que solo incluye algunas reglas sobre la edad y los antecedentes penales de un individuo. El modelo completo de ML predice que una persona será detenida nuevamente dentro de los dos años posteriores a su evaluación si ha cometido tres o más delitos anteriores, tiene entre 18 y 20 años y es hombre o tiene entre 21 y 23 años y ha cometido dos o tres delitos anteriores. Este conjunto de pautas es tan preciso como el modelo de caja negra COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) ampliamente utilizado (y patentado), que se utiliza en el condado de Broward, Florida. Consulte la sección sobre sesgo algorítmico para familiarizarse con COMPAS.

Caso de estudio: Guía de los Institutos Nacionales de Estándares y Tecnología de EE. UU. (NIST) sobre la explicabilidad de la IA

El NIST de EE. UU. ha emitido una guía sobre la explicabilidad de la IA que podría ser parte de los sistemas de evaluación de impacto. El borrador de las directrices del NIST sugiere cuatro principios para la explicación de las herramientas de evaluación de los sistemas de toma de decisiones automatizadas (Ads) sensibles a la audiencia y orientados a un propósito: (1) los sistemas ofrecen evidencia o razones adjuntas para todos los resultados; (2) los sistemas proporcionan explicaciones que son comprensibles para los usuarios individuales; (3) la explicación refleja correctamente el proceso del sistema para generar el resultado; y (4) el sistema solo funciona en condiciones para las que fue diseñado o cuando el sistema alcanza suficiente confianza en su resultado. Estos cuatro principios dan forma a los tipos de explicaciones necesarias para garantizar la confianza en los sistemas algorítmicos de toma de decisiones, como las explicaciones para el beneficio de la persona usuaria, para la aceptación social, para fines regulatorios y de cumplimiento, para el desarrollo del sistema y para el beneficio de la persona propietaria.

Fuente: NIST (2020). Four Principles of Explainable Artificial Intelligence, disponible en: https://www.nist.gov/system/files/documents/2020/08/17/NIST%20Explainable%20AI%20Draft%20NISTIR8312%20%281%29.pdf

4. El principio del humano en el circuito

Al darse cuenta de que muchos sistemas de IA son cajas negras y propensos a sesgos, los operadores judiciales comenzarán a abordar cuestiones relativas a la medida en que los seres humanos pueden o deben depender de la IA. ¿Deberían los seres humanos supervisar o aprobar ciertos resultados y decisiones recomendados por la IA antes de que se implementen? ¿Quién es responsable de las fallas o la piratería de las tecnologías basadas en IA? Se presentarán disputas sobre la incapacidad de las partes para comprender o gestionar completamente ciertas operaciones impulsadas por IA, así como disputas sobre lo que es justo en ADM.

Para la eficiencia y seguridad de las aplicaciones impulsadas por IA, los operadores judiciales deben asegurarse de que siempre haya un "humano en el circuito", es decir, que la IA nunca reemplace completamente a los humanos para que los profesionales adecuadamente capacitados validen las decisiones de IA. La IA es tan buena como los datos, el capital humano y la experiencia del equipo interdisciplinario involucrado en el desarrollo de la solución de IA. Un marco adecuado de IA y gobernanza de datos debe definir las responsabilidades respectivas de todas las partes interesadas, incluidas las partes interesadas del poder judicial. Debe establecer las condiciones y garantías necesarias para proteger los derechos humanos mientras se trabaja por el interés colectivo. Esto podría hacerse mediante

la certificación pública de los sistemas de IA que garantizarían la calidad de los datos y los algoritmos para evitar la profundización de las desigualdades existentes. La certificación pública de las aplicaciones de IA generaría confianza pública y permitiría a los usuarios dar su consentimiento informado.[46]

Por lo tanto, es importante poder medir el nivel de riesgo y el impacto de los diferentes sistemas de IA que podrían implementarse en el sistema de justicia. En este sentido, es importante determinar el requisito de supervisión humana, en función del caso de uso, su sensibilidad, la complejidad y opacidad del algoritmo y el posible impacto en los derechos humanos.[47] Como ejemplo, un jugador de ajedrez de IA con bajo riesgo solo podría necesitar una simple autoevaluación, educación del usuario y supervisión interna. Sin embargo, un cirujano de IA con alto riesgo podría exigir evaluaciones revisadas por pares, registros públicos, intervenciones humanas significativas, capacitación periódica y evaluación externa.

El Modelo de Marco de Gobernanza de Inteligencia Artificial, Segunda edición desarrollado por el Gobierno de Singapur (véase la Figura 8 a continuación) describe tres enfoques generales para la supervisión humana de los sistemas de IA: (i) humano-en-el-circuito, (ii) humano-fuera-del-circuito, y (iii) humano-dentro-del-circuito. La medida en que se necesita supervisión humana depende de los objetivos del sistema de IA y de una evaluación de riesgos, como se ilustra en los ejemplos a continuación.

- **El término "humano en el circuito" (HITL)** se refiere a un proceso en el que un sistema de IA es monitoreado de cerca por un humano, quien es responsable de tomar todas las decisiones finales. Esto es especialmente importante en campos como la atención médica, donde la IA puede proporcionar un apoyo invaluable para hacer recomendaciones para el tratamiento del cáncer, la terapia de la sepsis, la planificación quirúrgica y más. Si bien las herramientas de IA pueden ayudar a los proveedores de atención médica a tomar decisiones informadas de manera rápida y precisa, la responsabilidad final de la atención al paciente siempre recae en el experto humano.
- **El término "humano fuera del circuito"** se refiere a la ausencia de supervisión humana en las decisiones tomadas por el sistema de IA. Esto significa que el sistema de IA tiene completo control y no hay posibilidad de intervención humana. Un ejemplo de esto sería un sistema de ciberseguridad impulsado por IA que pueda detectar y corregir vulnerabilidades del sistema sin la necesidad de la participación humana. Mayhem, el sistema ganador del Gran Desafío Cibernético 2016 de la Agencia de Proyectos de Investigación Avanzada de Defensa (DARPA), es un sistema innovador que escanea

46 Stankovich M. (2021). Regulating AI and Big Data Deployment in Healthcare: Proposing Robust and Sustainable Solutions for Developing Countries 'Governments, disponible en: https://www.dai.com/uploads/regulating-ai-cda.pdf

47 Según el Grupo de expertos de alto nivel de la Comisión Europea sobre IA, "HITL se refiere a la capacidad de intervención humana en cada ciclo de decisión del sistema, que en muchos casos no es posible ni deseable. HOTL se refiere a la capacidad de intervención humana durante el ciclo de diseño del sistema y el monitoreo de la operación del sistema. HIC se refiere a la capacidad de supervisar la actividad global del sistema de IA (incluidas sus repercusiones económicas, sociales, jurídicas y éticas más amplias) y a la capacidad de decidir cuándo y cómo utilizar el sistema en una situación concreta", véase: https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines/1. Véase también el Modelo de Marco de Gobernanza de IA de Singapur, disponible en: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf

constantemente cualquier nueva vulnerabilidad que pueda explotarse por piratas informáticos. Cuando Mayhem detecta un nuevo error, genera automáticamente código para proteger el software de esta vulnerabilidad. Este sistema es un experto en analítica prescriptiva, lo que significa que puede detectar e interactuar con máquinas sin intervención humana. Esto contrasta con los sistemas tradicionales de detección de intrusiones que dependen de la intervención humana para anticiparse a los ciberataques.

- **El término "humano dentro del circuito"** se refiere a la participación de humanos en roles de supervisión en los que tienen la capacidad de tomar el control cuando los modelos de IA se encuentran con situaciones inesperadas o indeseables. Una forma efectiva de entender esto es a través de un sistema de navegación GPS. El sistema GPS planifica la ruta del punto A al B y ofrece varias opciones basadas en parámetros como la distancia más corta, el tiempo más corto o evitar las carreteras con peaje. Sin embargo, durante la navegación, el conductor aún puede hacerse cargo del GPS y modificar los parámetros de navegación en caso de congestión inesperada de la carretera.

Figura 8. Nivel de participación humana en el despliegue de IA

Fuente: IMDA, Singapur

Cabe señalar que el principio HITL tiene sus limitaciones debido al sesgo de automatización analizado en el Módulo 3, cuando los humanos están más predispuestos a las decisiones simples tomadas por algoritmos, especialmente en los casos en que hay un efecto de caja negra y los humanos podrían no ser capaces de entender por qué se tomó esta decisión.

5. ¿Por qué es importante la ciberseguridad en el contexto de la IA?

La ciberseguridad es la gestión de riesgos para la confidencialidad, integridad o disponibilidad de datos y sistemas. Es un tema fundamental para cualquier tecnología. Los procesos/algoritmos de IA procesan inherentemente grandes conjuntos de datos y con frecuencia producen resultados con consecuencias tanto virtuales como tangibles. Además de las amenazas tradicionales, se han identificado vulnerabilidades exclusivas de la IA, que incluyen:

- Envenenamiento de datos durante la etapa de entrenamiento[48]
- Ataques de entrada que manipulan los datos para alterar la salida[49]

Los ciberataques siguen aumentando en frecuencia, sofisticación y costos. En 2022, las empresas necesitan un promedio de 207 días para detectar un incidente de seguridad y 70 días para contenerlo. A medida que las empresas continúan implementando rápidamente la tecnología en toda la cadena de valor, el riesgo de interrupción del negocio asume un papel central. En casa, los dispositivos integrados de Internet de las cosas (IoT) siguen planteando riesgos significativos, y el trabajo remoto introduce una complicada combinación de vulnerabilidades. Los actores malintencionados pueden comprometer los sistemas de IA para lograr diversos objetivos, como causar daños, evadir la detección o degradar la fe en un sistema.[50]

En comparación con los sistemas tradicionales, los sistemas impulsados por IA presentan características únicas que pueden ser vulnerables a los ciberataques de formas no tradicionales. Por ejemplo, los atacantes pueden comprometer un conjunto de datos de entrenamiento para que el "aprendizaje" resultante del sistema no sea el previsto. Este tipo de ataque se llama envenenamiento de datos y aprovecha el proceso de desarrollo único de la IA, que es el uso de datos de gran tamaño. Por lo tanto, es importante proporcionar protección adicional a los sistemas de IA. El aumento de las capacidades de aprendizaje en las tecnologías de IA, como el aprendizaje profundo y el aprendizaje por refuerzo, tiene un impacto significativo en la ciberseguridad y permite acciones delictivas de manera más eficiente.[51] Por lo tanto, la protección de los sistemas de IA debe analizarse cuidadosamente e identificarse las posibles vulnerabilidades para poder implementar medidas de seguridad sólidas para (a) protegerse contra los ataques, pero también (b) detectar los ataques lo antes posible para mitigar los riesgos y daños significativos.

48 Poremba S. (2021). Data Poisoning: When Attackers Turn AI and ML Against You, disponible en: https://securityintelligence.com/articles/data-poisoning-ai-and-machine-learning/

49 Comiter M. (2019). Attacking Artificial Intelligence: AI 's Security Vulnerability and What Policymakers Can Do About It, disponible en: https://www.belfercenter.org/publication/AttackingAI

50 Ibid.

51 Kaloudi N., Li J. (2020). The AI-Based Cyber Threat Landscape: A Survey. ACM Computing Surveys (CSUR), 53, 1–34, disponible en: https://www.researchgate.net/publication/339081899_The_AI-Based_Cyber_Threat_Landscape_A_Survey.

- Los ciberataques a los sistemas de IA ocurren en tres fases diferentes de desarrollo de IA: 1) preparación de datos, 2) entrenamiento de modelos e 3) implementación de modelos: [52]
- Durante la preparación de datos, los atacantes pueden apuntar a componentes o bibliotecas comunes de preparación de datos, u obtener acceso no autorizado a la canalización de procesamiento de datos con fines de manipulación.
- Durante la fase de formación, los atacantes pueden añadir, eliminar o cambiar los datos de entrenamiento (envenenamiento de datos). Al hacer esto, los atacantes influyen en el modelo resultante.
- Los atacantes que tienen acceso a los modelos pueden introducir cambios en los pesos y algoritmos en la etapa de implementación del modelo (manipulación del modelo).[53]

Regulación de la ciberseguridad

La regulación de la ciberseguridad consiste en directivas que protegen la tecnología de la información y los sistemas informáticos para obligar a las entidades del sector privado y público a proteger sus sistemas de información y datos de ciberataques como virus, gusanos, troyanos, phishing, ataques de denegación de servicio (DOS), acceso no autorizado (robo de propiedad intelectual o información confidencial) y ataques al sistema de control[54].

Teniendo esto en cuenta, es extremadamente importante que los operadores judiciales tengan en cuenta las diferentes leyes y regulaciones de ciberseguridad y evalúen cómo la IA puede afectar estas regulaciones. Por ejemplo, las redes inteligentes que utilizan sistemas de IA mejorarán significativamente la gestión del consumo y la distribución de energía en beneficio de los consumidores, los proveedores de electricidad y los operadores de la red. No obstante, las operaciones y servicios mejorados expondrán a toda la red energética a nuevas dificultades en la seguridad del sistema de comunicación e información. Las vulnerabilidades de las redes de comunicación y los sistemas de información podrían explotarse por razones financieras o políticas para cortar la energía a áreas amplias o para lanzar ciberataques contra las unidades productoras de energía. La IA se puede utilizar en campañas de mala información y desinformación que podrían utilizarse para apagar Internet y restringir el acceso a la información.[55] El siguiente cuadro describe los peligros asociados con los ejemplos antagónicos utilizados por los modelos de ML.

52 Gartner (2020). Artificial Intelligence Under Attack: How to Identify and Mitigate Threats to Machine Learning, disponible en: https://www.gartner.com/en/documents/3989271; Wolff J. (2020). How to Improve Cybersecurity for Artificial Intelligence, disponible en: https://www.brookings.edu/articles/how-to-improve-cybersecurity-for-artificial-intelligence/

53 Ibid.

54 Ley de Ciberseguridad de la UE (2019), disponible en: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32019R0881&qid=1694014957942. Véase también: https://web.archive.org/web/20100613183200/http://www.privacyrights.org/ar/ChronDataBreaches.htm

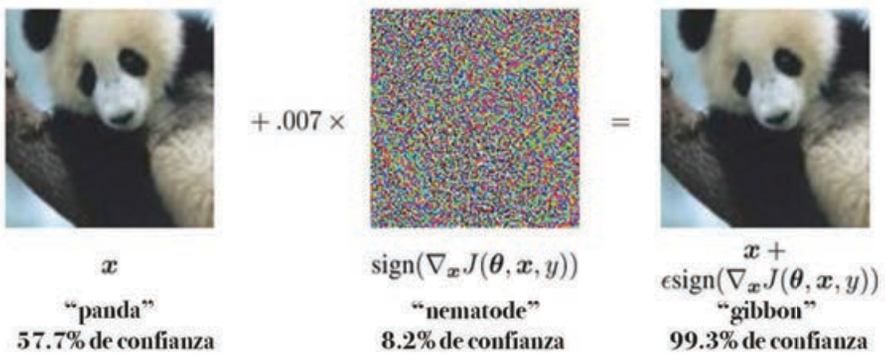
55 ENISA (2016). Inventario, análisis y recomendaciones sobre la protección de las IIC, disponible en: https://www.enisa.europa.eu/publications/stocktaking-analysis-and-recommendations-on-the-protection-of-ciis.

En profundidad: los peligros asociados con los ejemplos antagónicos utilizados por los modelos de ML

Los ejemplos antagónicos son entradas utilizadas por los modelos de ML que son generadas a propósito por un atacante para hacer que el modelo se equivoque mientras exhibe un alto nivel de confianza. Debido a que muchos modelos de ML, incluso las redes neuronales de vanguardia[56], son susceptibles a instancias antagónicas, esto puede representar una grave amenaza para la seguridad y la robustez de la IA.

Los ejemplos pueden pasar desapercibidos. La imagen de un panda a continuación ha sufrido una pequeña perturbación indetectable, o "entrada antagónica" insertada. Se pretende engañar al algoritmo de clasificación de imágenes. Esto ha dado como resultado que el computador tenga un nivel de confianza del 99,3 % en la clasificación del panda como gibón.

Se pueden producir ejemplos antagónicos imprimiendo una imagen en papel normal y tomando una foto de ella con un teléfono inteligente con una resolución típica. Un sticker antagónico en una señal de parada podría engañar a un automóvil autónomo para que piense que es una señal de "ceder el paso" o cualquier otra señal.[57]



Fuente: OCDE, AI in society, disponible en: https://www.oecd-ilibrary.org/science-and-technology/artificial-intelligence-in-society_eedfee77-en

Las debilidades de estos sistemas de IA contra ejemplos contradictorios tienen efectos perjudiciales en la seguridad de los sistemas de IA. La adopción de sistemas críticos como los utilizados en el transporte autónomo, las imágenes médicas y la seguridad y vigilancia podría sufrir seriamente por la existencia de casos en los que las perturbaciones sutiles pero específicas llevan a los modelos a sufrir engaños en grandes errores de cálculo y decisiones incorrectas.

56 Las redes neuronales son un tipo de técnica de ML que permite a los computadores aprender a realizar tareas analizando ejemplos de entrenamiento. Por lo general, estos ejemplos están preetiquetados. Por ejemplo, un sistema de reconocimiento de objetos puede recibir miles de imágenes etiquetadas de objetos como automóviles, casas y tazas de café. A través del análisis, puede identificar patrones en las imágenes que correspondan a las etiquetas específicas. Una red neuronal está diseñada para parecerse a la estructura del cerebro humano, con miles o millones de nodos de procesamiento interconectados. Estos nodos suelen estar organizados en capas y los datos fluyen a través de ellos en una sola dirección, lo que los convierte en "alimentación prospectiva". Cada nodo recibe datos de los nodos de la capa inferior y envía datos a los nodos de la capa superior. Definición proporcionada en Hardesty L. (2017). Explained neural networks, disponible en: https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

57 Goodfellow I. J., Shlens J., Szegedy (2015). Explicar y aprovechar ejemplos contradictorios. Conferencia Internacional sobre Representación del Aprendizaje, disponible en: https://arxiv.org/pdf/1412.6572.pdf; Kurakin A., Goodfellow I., Bengio S. (2017). Ejemplos antagónicos en el mundo físico. Taller ICLR, disponible en: https://arxiv.org/abs/1607.02533

6. Actividades

Estas actividades grupales tienen como objetivo alentar a las personas participantes de la capacitación a analizar y debatir varias preguntas pertinentes relacionadas con la IA y sus componentes básicos, y los riesgos asociados con el despliegue de la IA en el poder judicial.

Actividad 1: Tiempo de debate

Debata estas preguntas con otras personas participantes de la capacitación:

- ¿Cómo puede una persona acusada impugnar legítimamente la lógica de un algoritmo si el código fuente y (si corresponde) los datos de entrenamiento o los conjuntos de datos que se requerirán para reproducir los resultados no se ponen a su disposición?
- ¿Qué información se debe proporcionar a la persona acusada para impugnar la lógica de un algoritmo?
- ¿Es suficiente que tenga acceso simplemente a las entradas y salidas generadas por el algoritmo?
- ¿Debe la persona acusada recibir información sobre el margen de error de los algoritmos utilizados?

Actividad 2: Tiempo de debate

Debata estas preguntas con otras personas participantes de la capacitación:

- ¿Cómo pueden los tribunales hacer cumplir el debido proceso legal si el algoritmo implementa el aprendizaje automático y nadie, ni siquiera el desarrollador, entiende completamente el "análisis" de ML?
- ¿Cómo evaluarán los tribunales la precisión de los algoritmos, particularmente cuando pronostican el comportamiento humano futuro?

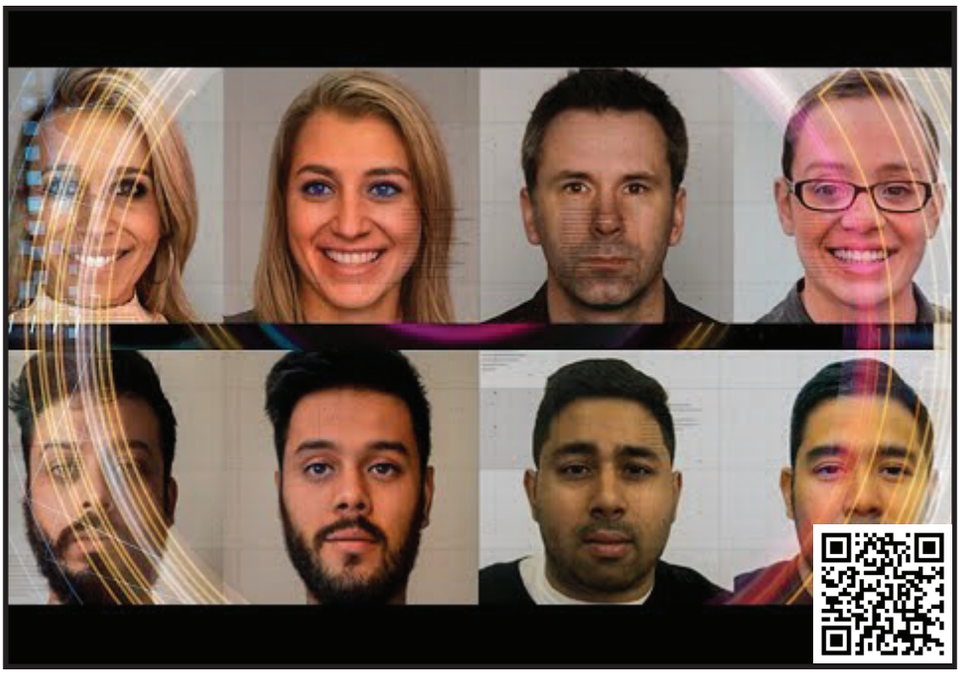
Actividad 3: Tiempo de debate

Debata estas preguntas con otras personas participantes de la capacitación:

- ¿Qué pasa si los algoritmos se han entrenado con conjuntos de datos anteriores que no incluyen la jurisprudencia más reciente?
- ¿Cuál es el régimen para la admisibilidad de las pruebas recopiladas con la ayuda de algoritmos, especialmente por parte de los investigadores policiales?
- ¿Se puede considerar que este acopio es irregular o injusto?
- ¿Se han recopilado los datos de conformidad con las leyes de protección de datos y, de no ser así, cómo debe tratarse el algoritmo?

Actividad 4: Tiempo de debate

Las personas participantes de la capacitación ven el video y debaten los diferentes impactos sociales del sesgo de la IA.



Fuente: BBC, https://youtu.be/b4UyT85H3Hg

Actividad 5: Los participantes de la capacitación debaten los siguientes temas relacionados con la aplicación de la IA en las operaciones judiciales.

A menudo, los modelos de IA no pueden proporcionar justificaciones comprensibles para sus decisiones o recomendaciones. Muchos algoritmos de IA "aprenden por sí mismos", es decir, ML de autoaprendizaje [También, lea y refiérase al principio humano en el circuito en el Módulo 4]. Intente responder las siguientes preguntas mientras habla con otras personas participantes de la capacitación:

- ¿Cómo afecta su capacidad para comprender o investigar los resultados de un modelo de IA a su valor probatorio en los procedimientos de litigio?
- ¿Qué responsabilidades legales y sociales debemos dar a los algoritmos protegidos detrás de la "imparcialidad" derivada de los datos estadísticos?
- ¿Quién es responsable cuando la IA se equivoca?

Existe un gran debate sobre quién, entre los diversos participantes y actores a lo largo del ciclo de vida de diseño, desarrollo e implementación de la IA y los sistemas autónomos, debe ser responsable frente a cualquier daño que pueda causarse. Un ecosistema complejo de IA y la multiplicidad de actores dificultan la determinación de quién puede considerarse responsable del daño causado al (a los) reclamante(s), ya que el daño puede ser el resultado de una serie de causas entrelazadas por múltiples actores.

- ¿Las capacidades de autonomía y autoaprendizaje alterarían la cadena de responsabilidad del productor o desarrollador como la "máquina impulsada por IA o automatizada de otro modo que, después de considerar ciertos datos, ha evolucionado con el tiempo a través de sus capacidades de autoaprendizaje habilitadas por ML y/o técnicas de aprendizaje profundo tomando una decisión autónoma y causando daño a la vida, la salud o la propiedad de un ser humano"?
- ¿Cómo afectarán las capacidades de los sistemas de ML no supervisados a los problemas de responsabilidad? Por ejemplo, un desafío es la dependencia de datos externos: cuando dichos datos se suministran de fuentes externas, probar tanto su carácter defectuoso como un vínculo causal con la lesión o el daño sufrido podría ser muy difícil.
- ¿"Insertar una capa de código inescrutable, poco intuitivo y estadísticamente derivado entre un tomador de decisiones humano y las consecuencias de esa decisión en la IA interrumpe nuestra comprensión típica de la responsabilidad por las elecciones que salen mal"? ¿O debería el productor o programador prever la posible pérdida o daño incluso cuando puede ser difícil anticipar, particularmente en circunstancias inusuales, las acciones de un sistema autónomo? Estas preguntas se volverán más importantes a medida que los sistemas de IA tomen decisiones cada vez más autónomas.
- ¿Qué niveles de incertidumbre en los resultados del modelo de ML aceptarán los tribunales y en qué condiciones? ¿Cómo se relacionan varios niveles de certeza del modelo de ML con varios estándares de evidencia? (Es decir, ¿cuándo el grado de correlación X, Y o Z [menos X %, Y % o Z % de incertidumbre] equivale a algún estándar legal de prueba [como "claro y convincente", "preponderancia de evidencia" o "más allá de una duda razonable"]?) ¿Será este un problema dejado a la discreción de los tribunales o jueces individuales? ¿O se desarrollarán e implementarán normas nacionales o regionales? ¿Deben ser estos requisitos rígidos o flexibles?
- La mayoría de los procesos de aprendizaje automático son iterativos y de autoaprendizaje, ya que ajustan las fórmulas y la precisión mediante el procesamiento de nuevos datos. ¿Cree que tales aplicaciones de ML, ya que cambian constantemente, podrían necesitar volverse a someter a un proceso judicial de forma continua si se usan como evidencia?

Fuente: AAAS, Artificial Intelligence and the Courts: Materials for Judges, disponible en: https://www.aaas.org/ai2/projects/law/judicialpapers.

7. Recursos

1. AccessNow (2018). Declaración de Toronto sobre la protección de los derechos a la igualdad y la no discriminación en los sistemas de aprendizaje automático, disponible en: https://www.accessnow.org/press-release/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/
2. AccessNow (2018). Declaración de Toronto sobre la protección de los derechos a la igualdad y la no discriminación en los sistemas de aprendizaje automático, disponible en: https://www.accessnow.org/press-release/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/
3. Amnistía Internacional (2017). Inteligencia artificial para el bien, disponible en: https://www.amnesty.org/en/latest/news/2017/06/artificial-intelligence-for-good
4. Amnistía Internacional (2021). Xenophobic Machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal, disponible en: Xenophobic machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal - Amnistía Internacional
5. Fuente: Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, disponible en: https://ssrn.com/abstract=4162985
6. Buolamwini J., Gebru T. (2018). Sombras de género: disparidades interseccionales de precisión en la clasificación comercial de género. Actas de la 1.ª Conferencia sobre Equidad, Rendición de Cuentas y Transparencia, PMLR, 81, 77–91, disponibles en: https://proceedings.mlr.press/v81/buolamwini18a.html
7. Burgess M. (2023). El agujero de seguridad en el corazón de ChatGPT y Bing. Disponible en:https://www.wired.co.uk/article/chatgpt-prompt-injection-attack-security
8. Burrell J. (2015). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms disponible en: https://ssrn.com/abstract=2660674 o http://dx.doi.org/10.2139/ssrn.2660674
9. Conn A. (2017). Artificial Intelligence: The Challenge to Keep It Safe., disponible en: https://futureoflife.org/ai/safety-principle/European Union Agency for Fundamental Rights (2019), disponible en: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2019-data-quality-and-ai_en.pdf
10. IEEE (2019). Diseño éticamente alineado. Una visión para priorizar el bienestar humano con sistemas autónomos e inteligentes. Primera edición, disponible en: https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e.pdf
11. Oficina del Comisionado Internacional, Explaining decision made with AI, disponible en: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/
12. McGregor L., Murray D., Ng V. (2019). International Human Rights Law as a Framework for Algorithmic Accountability, International & Comparative Law Quarterly, 68(2), 309–343, disponible en: www.cambridge.org/core/journals/international-and-comparative-law-quarterly/article/international-human-rights-law-as-a-framework-for-algorithmic-accountability/1D6D0A456B36BA7512A6AFF17F16E9B6
13. Consejo Nacional de Ciencia y Tecnología: Comité de Tecnología (2016). Preparing for the Future of Artificial Intelligence. Washington, D.C.: Oficina Ejecutiva del Presidente,

2016. https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf

14. Noble S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism, New York University Press.
15. Obermeyer Z., Powers B., Vogeli C., Mullainathan S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations, Science, 366(6464), 447–453, disponible en: https://www.science.org/doi/10.1126/science.aax2342
16. OCDE (2022). Marco para la clasificación de los sistemas de IA, disponible en: https://www.oecd.org/publications/oecd-framework-for-the-classification-of-ai-systems-cb6d9eca-en.htm.
17. O'Neil C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Nueva York: Crown.
18. Stanford (2022). Artificial Intelligence Index Report, available at: 2022-AI-Index-Report_Master.pdf (stanford.edu)
19. The Alan Turing Institute, Human Rights, Democracy, and the Rule of Law Assurance Framework for AI Systems: A proposal prepared for the Council of Europe 's Ad hoc Committee on Artificial Intelligence, disponible en: https://www.turing.ac.uk/news/publications/ai-human-rights-democracy-and-rule-law-primer-prepared-council-europe
20. The Royal Society (2012). Machine Learning: The Power and Promise of Computers that Learn by Example, disponible en: https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf 16
21. UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, available at: https://unesdoc.unesco.org/ark:/48223/pf0000380455
22. Ward J. (2019). 10 Things Judges Should Know About AI, Judicature, 103(1), disponible en: https://judicature.duke.edu/articles/10-things-judges-should-know-about-ai.
23. Weinberger D. (2017). Our Machines Now Have Knowledge We'll Never Understand, disponible en: https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/
24. Wong A. (2020). The Laws and Regulation of AI and Autonomous Systems. En: Strous L., Johnson R., Grier D. A., Swade D. (eds) Unimagined Futures – ICT Opportunities and Challenges, IFIP Advances in Information and Communication Technology(), 555, disponible en: https://link.springer.com/chapter/10.1007/978-3-030-64246-4_4
25. Wong A. (2021). Ethics and Regulation of Artificial Intelligence. En: Mercier-Laurent E., Kayalica M.Ö., Owoc M.L. (eds) Artificial Intelligence for Knowledge Management, AI4KM, IFIP Advances in Information and Communication Technology, 614, disponible en: https://www.researchgate.net/publication/352477342_Ethics_and_Regulation_of_Artificial_Intelligence
26. Wong A. (2023). Generative AI: The Global debate and controversies on use of copyrighted content as training data, disponible en: https://unctad.org/news/cstd-dialogue-anthony-wong

Módulo 2

Adopción de IA en el poder judicial

El módulo dos aborda la adopción de la IA en el poder judicial. Presenta las distintas aplicaciones de la IA en el poder judicial, como el descubrimiento electrónico y la revisión de documentos, el uso de la IA generativa para ayudar en la redacción de documentos, el análisis predictivo, las herramientas de evaluación de riesgos, la resolución de litigios, el reconocimiento del lenguaje, el expediente digital y la gestión de casos. El Módulo luego destaca estudios de casos sobre el despliegue de la IA en el poder judicial, debatiendo algunas de las oportunidades y desafíos que enfrentan los sistemas judiciales en todo el mundo en el uso de la IA.

¿Qué va a aprender?

Después de completar este módulo, las personas participantes podrán:

- Comprender las diferentes aplicaciones de la IA en el poder judicial;
- Comprender los desafíos y oportunidades relacionados con el despliegue de sistemas de IA en el poder judicial a través de los casos de estudio presentados en el módulo.

1. ¿Cuáles son las aplicaciones de la IA en el poder judicial?

Abogados, bufetes de abogados, tribunales y agencias gubernamentales están utilizando la IA para diferentes propósitos. Por ejemplo, los abogados están utilizando la IA para la investigación jurídica y para encontrar precedentes relevantes para fortalecer sus argumentos. Los bufetes de abogados lo utilizan para pronosticar los resultados de los casos, evaluar las posibilidades de éxito y asesorar a los clientes con respecto a los procedimientos legales Los abogados también han utilizado la IA para pronosticar cómo se pronunciarían los jueces sobre diversos temas. De manera similar, las entidades gubernamentales están utilizando la IA para evaluar la probabilidad de éxito en la búsqueda de cursos de acción particulares contra individuos y empresas, como en casos relacionados con impuestos.

En Buenos Aires, Argentina, los fiscales utilizan sistemas de inteligencia artificial para redactar fallos judiciales[58]. El Tribunal de Internet de Hangzhou ha implementado un sistema de análisis de pruebas que utiliza tecnologías de vanguardia como blockchain, IA, big data y computación en la nube. Este sistema analiza y compara todas las pruebas presentadas por ambas partes, transformándolas en una lista de evidencias y pruebas pertinentes. Luego, la información se divide y clasifica antes de presentarse visualmente al juez humano para su consideración.[59] En México, los tribunales pueden utilizar la IA para decidir sobre si alguien tiene derecho a una forma de seguridad social o no. Un programa llamado Expertius basa sus cálculos en información sobre reclamaciones pasadas, resultados de las reclamaciones, registros de audiencias y sentencias.[60]

Otro ejemplo es el sistema de justicia colombiano, que está explorando formas de reducir la carga de trabajo de los jueces humanos. La Corte Constitucional colombiana está desarrollando actualmente un sistema de IA llamado PretorIA para ayudar en la selección de tutores. PretorIA no reemplaza a los humanos en este proceso, sino que agiliza la tarea al analizar las sentencias de tutela y proporciona información más refinada a los responsables de identificar a las personas que pueden ser seleccionadas como tutores.[61]

El impulso por una justicia eficiente en medio de las restricciones presupuestarias

Como ocurre con otros servicios al consumidor, se espera que los tribunales presten servicios judiciales modernos, digitales y con capacidad de respuesta, reduciendo al mismo tiempo la pendencia de asuntos en un contexto de crecientes restricciones presupuestarias. Los sistemas de justicia habilitados por IA prometen aumentar la calidad de los servicios al tiempo que reducen los gastos relacionados con las operaciones judiciales.[62]

58 Dejusticia (2021). Conoce nuestra Investigación sobre PretorIA, la tecnología que incorpora la Inteligencia Artificial a la Corte Constitucional, disponible en: https://www.dejusticia.org/conoce-nuestra-investigacion-sobre-pretoria-la-tecnologia-que-incorpora-la-inteligencia-artificial-a-la-corte-constitucional/

59 Xuan H. (2021). One-Click Access to Evidence Analysis Results. Hangzhou Internet Court Launches Intelligent Evidence Analysis System, China Courts Network, disponible en: https://www.chinacourt.org/article/detail/2019/12/id/4747683.shtml

60 Goretty C., Martinez B. (2012). La inteligencia artificial y su aplicación al campo del Derecho, Alegatos, 82, 827–846, disponible en: http://alegatos.azc.uam.mx/index.php/ra/article/viewFile/205/184

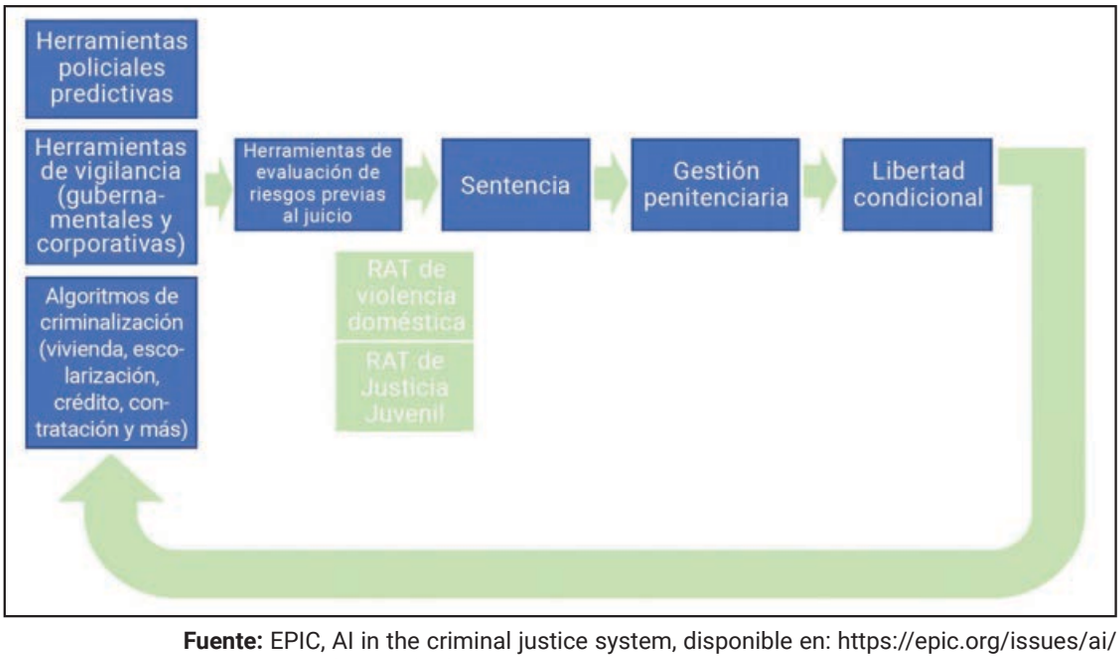
61 https://www.dejusticia.org/conoce-nuestra-investigacion-sobre-pretoria-la-tecnologia-que-incorpora-la-inteligencia-artificial-a-la-corte-constitucional/

62 Wu J. (2019). AI Goes to Court: The Growing Landscape of AI for Access to Justice, disponible en: https://medium.com/legal-design-and-innovation/ai-goes-to-court-the-growing-landscape-of-ai-for-access-to-justice-3f58aca4306f

Cuando se implementan con derechos humanos y salvaguardias éticas, los sistemas de IA pueden hacer que los procedimientos legales sean más accesibles para un grupo más amplio de personas, en múltiples idiomas y a menores costos. Por ejemplo, las estimaciones muestran que el uso de ML en el descubrimiento electrónico mediante la presentación de los documentos en grupos conceptuales puede aumentar la velocidad de revisión entre un 15 y un 20 por ciento. Esto supone un importante ahorro de costos.[63]

Por otro lado, el desarrollo y despliegue de IA en las operaciones judiciales puede afectar a los derechos fundamentales. Las tecnologías de IA contienen sesgos integrados (analizados en el Módulo 3), y a menudo son cajas negras (analizadas en el Módulo 1). Por lo tanto, el Estado de derecho y el mantenimiento de los derechos humanos deben seguir estando a la vanguardia de la administración de justicia.[64]

Figura 9. Un ciclo simplista de uso algorítmico en la justicia penal



Fuente: EPIC, AI in the criminal justice system, disponible en: https://epic.org/issues/ai/ai-in-the-criminal-justice-system/

La digitalización de los documentos judiciales en un primer paso esencial hacia el uso de la IA

La digitalización de los documentos judiciales ha permitido a los tribunales y otros operadores judiciales confiar en la asistencia de IA para las funciones administrativas. Los algoritmos de IA se utilizan cada vez más en el contexto de los sistemas de justicia civil y penal para apoyar la toma de decisiones humanas.[65] Los sistemas de IA se prueban para identificar patrones en la toma de decisiones judiciales complejas y predecir los resultados de las decisiones. A medida que los sistemas de IA recopilan y analizan grandes

63 Deloitte, Artificial intelligence and machine learning in e-discovery and beyond: Driving efficiencies in e-discovery using AI, disponible en: https://www2.deloitte.com/ch/en/pages/forensics/articles/AI-and-machine-learning-in-E-discovery.html 65 Dejusticia (2021). Conoce nuestra Investigación sobre PretorIA, la tecnología que incorpora la Inteligencia Artificial a la Corte Constitucional,

64 On the impact of AI on human rights when applied in the judicial systems, ver también UNESCO (2021). Global Toolkit for Judicial Actors: International legal standards on freedom of expression, access to information and safety of journalists, Module 5, p. 164, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000378755

65 Parlamento Europeo (2019). A governance framework for algorithmic accountability and transparency, disponible en: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262

cantidades de información, identifican patrones, predicen enfoques óptimos, detectan anomalías, clasifican problemas y redactan documentos, la promesa es que los sistemas judiciales serán más eficientes y podrán priorizar el tiempo y los recursos para garantizar la justicia oportuna.

En el sistema de justicia penal, se han desplegado modelos de IA para monitorear y reconocer a los acusados; apoyar las decisiones de sentencia y fianza; y apoyar la evaluación de la evidencia.[66] La Figura 9 ofrece una descripción simple del uso algorítmico de la IA en el sistema de justicia penal.

En el sistema de justicia civil, la IA se ha desplegado en procesos relacionados con la familia, la vivienda, la deuda, el empleo y el consumidor.[67] Los tribunales civiles recopilan cada vez más datos sobre la administración, los alegatos, el comportamiento de los litigantes y las decisiones. Lo anterior ofrece oportunidades para automatizar ciertas funciones judiciales, como la gestión de expedientes, la programación de audiencias y juicios, y la gestión de las funciones del jurado, lo que a su vez puede conducir a una mayor eficiencia.[68] Por ejemplo, la IA se utiliza para redactar previamente plantillas de juicios para jueces, hacer predicciones o recomendaciones de sentencias para fianzas, sentencias y cálculos financieros. También se utiliza para evaluar el resultado de los casos en función de las actividades pasadas de fiscales y jueces. Una herramienta de IA puede proporcionar información a un juez que tiene en cuenta una amplia cantidad de jurisprudencia y puede disminuir el tiempo de investigación en la preparación de decisiones.

Utilizando un algoritmo de IA creado por investigadores de la Universidad Católica de Lovaina (UCL), la Universidad de Sheffield y la Universidad de Pensilvania, se han anticipado las sentencias judiciales del Tribunal Europeo de Derechos Humanos con una precisión del 79 %.[69] El Dr. Nikolaos Aletras, quien dirigió el estudio en UCL Computer Science, explicó que "no vemos que la IA reemplace a jueces o abogados, pero creemos que les resultaría útil para identificar rápidamente patrones en casos que conducen a ciertos resultados. También podría ser una herramienta valiosa para resaltar qué casos tienen más probabilidades de constituir violaciones del Convenio Europeo de Derechos Humanos ".[70]

El desafío de que los sistemas de IA se perciban como más objetivos que los humanos

Sin embargo, dado el alto número de casos y la falta de recursos adecuados que afectan a la mayoría de los sistemas judiciales, existe el riesgo de que los jueces utilicen incorrectamente los sistemas de apoyo basados en IA para "delegar" decisiones a sistemas tecnológicos que no fueron diseñados para ese propósito pero que se perciben como más objetivos de lo que son. Para no poner en peligro el derecho a un juicio justo, se debe

66 Završnik A. (2020). Criminal justice, artificial intelligence systems, and human rights. Foro ERA. 20, 567-583, disponible en: https://doi.org/10.1007/s12027-020-00602-0.
67 Cabral J. E, Chavan A., Clarke T. M., Greacen J., Hough B. R., Rexer L., Ribadeneyra L., Zorza R. (2012). Using Technology to enhance access to justice, disponible en: http://jolt.law.harvard.edu/articles/pdf/v26/26HarvJLTech241.pdf
68 Martin A. (2010). Automated Debt-Collection Lawsuits Engulf Courts, disponible en: https://www.nytimes.com/2010/07/13/business/13collection.html
69 UCL (2016). AI predicts outcomes of human rights trials, disponible en: https://www.ucl.ac.uk/news/2016/oct/ai-predicts-outcomes-human-rights-trials
70 Ibid.

tener mucho cuidado en evaluar de qué son capaces estos dispositivos y en qué condiciones se pueden desplegar. Esto es especialmente cierto cuando dichos sistemas se utilizan para emitir decisiones de libertad condicional. En un sistema de justicia basado en algoritmos, los jueces no deben ser simples aplicadores de algoritmos, sino también sus evaluadores críticos. La siguiente tabla describe las principales implicaciones positivas y negativas del uso de ADM e IA en el sistema de justicia.

Tabla 3. Consecuencias positivas y negativas del uso de la IA en el sistema de justicia

	Consecuencias positivas	Consecuencias negativas
Excelencia judicial	Ofrece a los jueces un análisis rápido de una serie de casos y factores; Acelera la investigación y la redacción; La optimización de procesos, la reducción de costos, el aumento de la agilidad, las ganancias de productividad, la eliminación del trabajo mecánico y repetitivo aumentan la seguridad jurídica.	Incorporación de prejuicios raciales, de género/sexo y de otro tipo; Reduce la discrecionalidad judicial y el elemento humano en la toma de decisiones; Muy difícil de usar; Amenaza a la independencia judicial, sesgo automatizado; La elaboración de perfiles de los jueces puede afectar al derecho fundamental a la protección de los datos personales, puede crear presiones y afectar a la independencia judicial.
Privacidad y seguridad	Protocolos de seguridad automáticos y limpieza de datos, lo que permite una mayor precisión en las salidas de IA.	Hacking, filtraciones de datos.
Propiedad de los datos	Los datos agregados por sistemas de IA se pueden utilizar para identificar tendencias, brechas de servicio e innovación.[71]	Dependiendo de la propiedad del sistema, los socios del sector privado podrían tener acceso a los datos personales; Los datos agregados se pueden utilizar para identificar y discriminar a individuos o grupos; La regulación limitada de la propiedad de los datos limita la protección de los derechos y la reparación para las personas afectadas por los sistemas de IA.
Estado de derecho	Evita que un gran interés se apropie del sistema de justicia.	Puede invadir los derechos fundamentales como se debatió en el Módulo 4; Amenazas a la democracia como desinformación, mala información, fraudes, propaganda, falsedades, operaciones de influencia o manipulación de la opinión pública, principalmente en procesos electorales.
Acceso a la justicia	Puede identificar patrones de sesgo contra grupos vulnerables en la toma de decisiones y los servicios Puede hacer que los plazos judiciales sean más rápidos y predecibles.	No está disponible de manera uniforme para que las partes analicen los datos o respalden su caso derivado de problemas de infraestructura y acceso (electricidad, internet, hardware), La falta de capacitación de los operadores y asistentes judiciales podría afectar los resultados positivos que podría aportar la IA.

Adaptado de UNDP (2021) Emerging Technologies and Judicial Integrity Toolkit for Judges.

Fuente: https://www.undp.org/asia-pacific/emerging-technologies-and-judicial-integrity

71 IBM (2021). Data aggregation involves gathering a significant amount of information from a database and presenting it in a more manageable and inclusive format, disponible en: https://www.ibm.com/docs/en/tnpm/1.4.2?topic=data-aggregation

Figura 10. Aplicaciones clave de la IA en el poder judicial

Fuente: Autores.

Descubrimiento electrónico y revisión de documentos

Las herramientas de IA se utilizan en el poder judicial para identificar, clasificar y revisar (i) normas legales, retenciones legales y hallazgos fácticos; (ii) argumentos que explican conclusiones y explicaciones de razones, y (iii) consideraciones legales específicas y elementos probatorios.

El descubrimiento electrónico es la identificación, recopilación y producción de información almacenada electrónicamente (IAE) en respuesta a una solicitud de divulgación en un procedimiento judicial o investigación. La IAE puede consistir en correos electrónicos, documentos, presentaciones, bases de datos, archivos de audio y video y sitios web.[72]

Actividad: Piense en cómo la IA puede cambiar el proceso de descubrimiento y debátalo con otros participantes de la capacitación.

Cuestiones a considerar: ¿Cuáles serán los estándares para la admisibilidad de declaraciones u otras pruebas, o ideas generadas por IA y/o en las que confían (o rechazan) los humanos? ¿Cómo evaluaremos su credibilidad o autenticidad?

72 https://cdslegal.com/knowledge/the-basics-what-is-e-discovery/

El descubrimiento electrónico se basa en la agrupación, un ejemplo de ML no supervisado, donde los elementos "similares" (por ejemplo, documentos) se agrupan para que las personas usuarias puedan reconocer sus características similares y aprender sobre la composición del conjunto de datos. Las personas usuarias no tienen control sobre las dimensiones a lo largo de las cuales se define la "similitud" y no tienen que etiquetar ejemplos de elementos en cada grupo para entrenar el sistema. Sin embargo, el diseñador del sistema debe especificar las características a lo largo de las cuales se debe medir la similitud del elemento y el número de grupos.[73] Por ejemplo, si se le indica al sistema de ML que identifique cualquier información sobre tenis y béisbol en los archivos, el algoritmo también agrupará archivos que contienen información sobre todo tipo de deportes.[74] Del mismo modo, una búsqueda de "pequeño sobre marrón" o "grasa" agrupará información sobre todo lo relacionado con la corrupción.[75]

La **búsqueda de conceptos** es otro método de ML no supervisado utilizado en el descubrimiento electrónico, donde el computador aprende el contexto en el que se utilizan las palabras y modela las relaciones entre las palabras. Las personas usuarias pueden buscar por significado y no por términos individuales. Es probable que un documento que contenga palabras como "abogado", "contrato" o "causa civil" sea un documento legal. El uso de cualquiera de estas palabras puede llevar a la conclusión de que el tema del documento es legal.[76]

La **Revisión Asistida por Tecnología (TAR)** o codificación predictiva es una técnica de ML supervisada en la que los computadores aprenden a distinguir los documentos relevantes de los irrelevantes en función de la codificación realizada por revisores humanos, y luego clasifican los documentos no etiquetados sin ayuda.[77] Por ejemplo, CLAUDETTE (Automated CLAUseDETectEr) es un proyecto de investigación interdisciplinar con base en el Departamento de Derecho del Instituto Universitario Europeo y una plataforma para el análisis y la anotación automatizada de documentos legales, y la detección de anomalías.[78]

Además, las herramientas de IA se pueden implementar para anonimizar información personal, confidencial o privilegiada incluida en registros electrónicos. Esto puede ayudar al cumplimiento de la normativa de protección de datos.[79]

Actividad: ¿Cómo funciona CLAUDETTE?

El objetivo de CLAUDETTE es empoderar a los consumidores y a la sociedad civil mediante el desarrollo de herramientas para el usuario final que permitan a todos evaluar fácilmente la imparcialidad de los contratos de los consumidores y las regulaciones de privacidad antes de utilizar las plataformas de Internet. La tecnología se encuentra actualmente en fase experimental de laboratorio, y los participantes en la formación pueden acceder a ella aquí: http://claudette.eui.eu/demo

Los participantes de la capacitación ven el video (http://claudette.eui.eu/claudette.mp4) y analizan si tienen plataformas similares en sus respectivas jurisdicciones. ¿Cuáles son las ventajas y desventajas de la TAR?

73 EDRM (2021). The Use of Artificial Intelligence in eDiscovery, disponible en: https://edrm.net/download/152621/75 IBM (2021). Data aggregation involves gathering a significant amount of information from a database and presenting it in a more manageable

74 Deloitte. Artificial intelligence and machine learning in e-discovery and beyondnDriving efficiencies in e-discovery using AI, disponible en: https://www2.deloitte.com/ch/en/pages/forensics/articles/AI-and-machine-learning-in-E-discovery.html

75 Ibid.

76 EDRM (2021). The Use of Artificial Intelligence in eDiscovery, disponible en: https://edrm.net/download/152621/

77 Ibid.

78 EUI. CLAUDETTE, disponible en: http://claudette.eui.eu/about/index.html

79 EDRM (2021). The Use of Artificial Intelligence in eDiscovery, disponible en: https://edrm.net/download/152621/

Análisis predictivo y soporte ADM

Con frecuencia, los sistemas de IA se utilizan como herramientas de pronóstico. Analizan grandes cantidades de datos, incluidos datos históricos, para evaluar riesgos y predecir tendencias futuras utilizando algoritmos. Los datos de entrenamiento pueden contener antecedentes penales, registros de arrestos, estadísticas de delitos, registros de intervenciones policiales en ciertos vecindarios, publicaciones en redes sociales, datos de comunicaciones y registros de viajes. Los sistemas predictivos pueden ayudar a los jueces a tener un mejor conocimiento de las tendencias en la jurisprudencia y a anticipar cómo se situará una posible decisión en el contexto de la jurisprudencia.[80]

El análisis predictivo es la categoría general de herramientas y modelos estadísticos, por ejemplo, sistemas de aprendizaje automático, que utilizan y analizan datos históricos para crear predicciones sobre el futuro para guiar la toma de decisiones. Estas predicciones pueden ser de bajo riesgo (por ejemplo, qué película recomendar), de riesgo medio (qué solicitud de préstamo proponer aceptar) o de alto riesgo (qué acusado tiene más posibilidades de involucrarse en un comportamiento en particular).[81]

El desarrollo de aplicaciones de IA que pronostican cómo un tribunal resolverá una demanda, caso o acuerdo es una aplicación de IA de rápido crecimiento en el sector de la justicia. Por ejemplo, las tecnologías de IA ya se están utilizando para perfilar a las personas, identificar lugares como posibles sitios de actividad delictiva o señalar a futuros reincidentes.[82] Estas prácticas son muy controvertidas, como se detalla en los Módulos 3 y 4.

Un ejemplo de ello es el sistema EXPERTIUS en México, que asesora a jueces y funcionarios sobre si un demandante tiene el derecho a una pensión. El programa consta de tres módulos; primero, brinda a los jueces y empleados la oportunidad de comprender el proceso (el módulo tutorial); segundo, permite a las personas usuarias proporcionar pruebas en apoyo de su caso y asignar “pesos” a cada pieza de documentación de respaldo (el módulo inferencial); y tercero, permite a las personas usuarias calcular el monto de la pensión a la que tienen derecho en función de criterios socioeconómicos específicos (el módulo financiero).[83]

80 Comité de expertos en intermediarios de Internet (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Council of Europe Study, DGI/2017/12, disponible en: https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

81 AAAS, Artificial Intelligence and the Courts: Materials for Judges, disponible en: https://www.aaas.org/ai2/projects/law/judicialpapers/

82 RAND (2013). Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operation, disponible en: www.rand.org/content/dam/rand/pubs/research_reports/RR200/RR233/RAND_RR233.pdf.

83 Bell F., Bennett Moses L., Legg M., Silove J., Zalnieriute M. (2022). AI Decision-Making and the Courts: A Guide for Judges, Tribunal Members and Court Administrators, Australasian Institute of Judicial Administration, disponible en: https://ssrn.com/abstract=4162985

Caso de estudio: El caso del sistema australiano Split UP

Un grupo de expertos en IA y abogados han desarrollado el sistema Split-Up que se utiliza en los tribunales australianos de derecho de familia. El sistema Split-Up utiliza el razonamiento basado en reglas junto con las redes neuronales para anticipar los resultados de las controversias de propiedad en el divorcio y otros asuntos de derecho de familia.

El sistema de separación es utilizado por los jueces para apoyar su toma de decisiones al ayudarlos a identificar los bienes conyugales que deben incluirse en un acuerdo. El sistema ayuda al juez a determinar qué porcentaje del patrimonio común debe recibir cada parte en función de factores como las contribuciones, las fuentes de ingresos y las necesidades futuras. El sistema analiza 94 elementos clave utilizando técnicas estadísticas basadas en la arquitectura de redes neuronales. El juez puede entonces decretar una sentencia de propiedad basada en el análisis realizado por el algoritmo. El sistema también tiene como objetivo proporcionar justificaciones claras para sus decisiones.

Un desafío en términos de sesgo al usar sistemas como Split-Up es que los datos utilizados en este contexto (las disputas de divorcio generalmente están marcadas por desequilibrios de género y los datos históricos pueden presentar un patrón de discriminación) podrían ser percibidos como verdades reales por las máquinas. Los operadores judiciales deben ser conscientes de estos desafíos y riesgos que vienen con los sistemas de IA como Split-Up.

Fuente: Zeleznikow J., Stranieri A. (1995). The split-up system: integrating neural networks and rule-based reasoning in the legal domain, ICAIL '95: Proceedings of the 5th international conference on Artificial intelligence and law, 185–194, disponible en: https://dl.acm.org/doi/10.1145/222092.222235

Herramientas de evaluación de riesgos (predicción de riesgos, modelado de riesgos y puntuación social)

Cada vez más, se utilizan herramientas de evaluación de riesgos basadas en datos para anticipar la probabilidad de un comportamiento delictivo futuro. En varios países, estas tecnologías se están utilizando para ayudar a la toma de decisiones en el sistema de justicia penal, incluidas las sentencias relativas a la sentencia, la libertad bajo fianza y las limitaciones posteriores a la sentencia para aquellos que se considera que pueden cometer otros delitos. Estas herramientas aprovechan los datos históricos para evaluar la probabilidad de que una persona tenga un riesgo "alto", "medio" o "bajo" de no presentarse a sus citas judiciales o de volver a ser arrestada. El algoritmo considera factores como los antecedentes penales y la edad en el momento del arresto, y genera una puntuación que los jueces utilizan para decidir si mantener a alguien en la cárcel o liberarlo.[84]

Para evaluar el riesgo de reincidencia de una persona e identificar las áreas de intervención, se utilizan herramientas de evaluación de riesgos en varias fases del proceso legal. Por ejemplo, se utilizan evaluaciones de riesgos:

i) Antes del juicio para guiar las opciones sobre la libertad en espera de resolución o el encarcelamiento. Por los departamentos de libertad condicional y cumplimiento alternativo de la pena para determinar la cantidad apropiada de supervisión, que puede incluir monitoreo electrónico y confinamiento en el hogar.

84 Wykstra S. (2018). Bail reform, which could save millions of unconvicted people from jail, explained, disponible en: https://www.vox.com/future-perfect/2018/10/17/17955306/bail-reform-criminal-justice-inequality

ii) Como parte de los planes de reingreso y supervisión, los administradores de casos y los proveedores de tratamiento implementan evaluaciones de riesgos para identificar las necesidades del cliente y conectarlos con los servicios adecuados.[85]

Las técnicas de evaluación de riesgos, según sus proponentes, hacen que el sistema de justicia penal sea más equitativo.[86] Los defensores de tales sistemas argumentan que la IA podría sustituir la intuición y el sesgo de los jueces, particularmente el sesgo racial, con una puntuación de evaluación de riesgos que parece ser más "objetiva".[87]

Sin embargo, en la práctica, numerosos estudios han demostrado que estas herramientas podrían incorporar y amplificar los sesgos hacia las poblaciones marginadas y vulnerables. Varios derechos humanos pueden verse implicados en el uso de la IA en el sistema de justicia penal, incluidos los derechos a la igualdad y la no discriminación, la igualdad ante la ley, la seguridad y la libertad personales, el derecho a la privacidad, el derecho a una audiencia pública y justa, la imparcialidad procesal y la presunción de inocencia (consulte la Figura 11 que brinda una descripción general de cómo las herramientas de evaluación de riesgos de la justicia penal afectan los derechos humanos; para obtener ejemplos específicos, consulte el Módulo 4 de este Kit de herramientas).[88] Para ilustrar estos puntos, algunas herramientas de evaluación de riesgos se basan en datos de llamadas policiales, que pueden ser un indicador poco fiable de los patrones delictivos reales (en relación con los registros de arrestos). Estos datos a menudo se distorsionan aún más por prejuicios raciales, como se ve en el infame caso de Amy Cooper, quien llamó a la policía sobre un observador de aves negro por simplemente pedirle que atara a su perro en Central Park.[89] Es crucial entender que el hecho de que se haga una llamada para denunciar un delito, no significa necesariamente que realmente haya ocurrido un delito. Sin embargo, tales llamadas pueden usarse como puntos de datos en los sistemas de evaluación de riesgos para justificar el envío de policías a un vecindario en particular o incluso dirigirse a un individuo específico, creando así un circuito de retroalimentación donde las tecnologías basadas en datos legitiman la vigilancia discriminatoria.[90]

En el caso de Ewert vs. Canadá, el Tribunal Supremo de Canadá enfatizó que las herramientas de evaluación de riesgos que se crean y verifican utilizando datos de grupos mayoritarios pueden no ser precisas para predecir las mismas características en grupos minoritarios.[91]

85 Comité de expertos en intermediarios de Internet (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Council of Europe Study, DGI/2017/12, disponible en: https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

86 Hao K., Stray J. (2019). ¿Puede hacer que la IA sea más justa que un juez? Juegue nuestro juego de algoritmos para tribunales, disponible en: https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/.

87 Wykstra S. (2018). Bail reform, which could save millions of unconvicted people from jail, explained, disponible en: https://www.vox.com/future-perfect/2018/10/17/17955306/bail-reform-criminal-justice-inequality

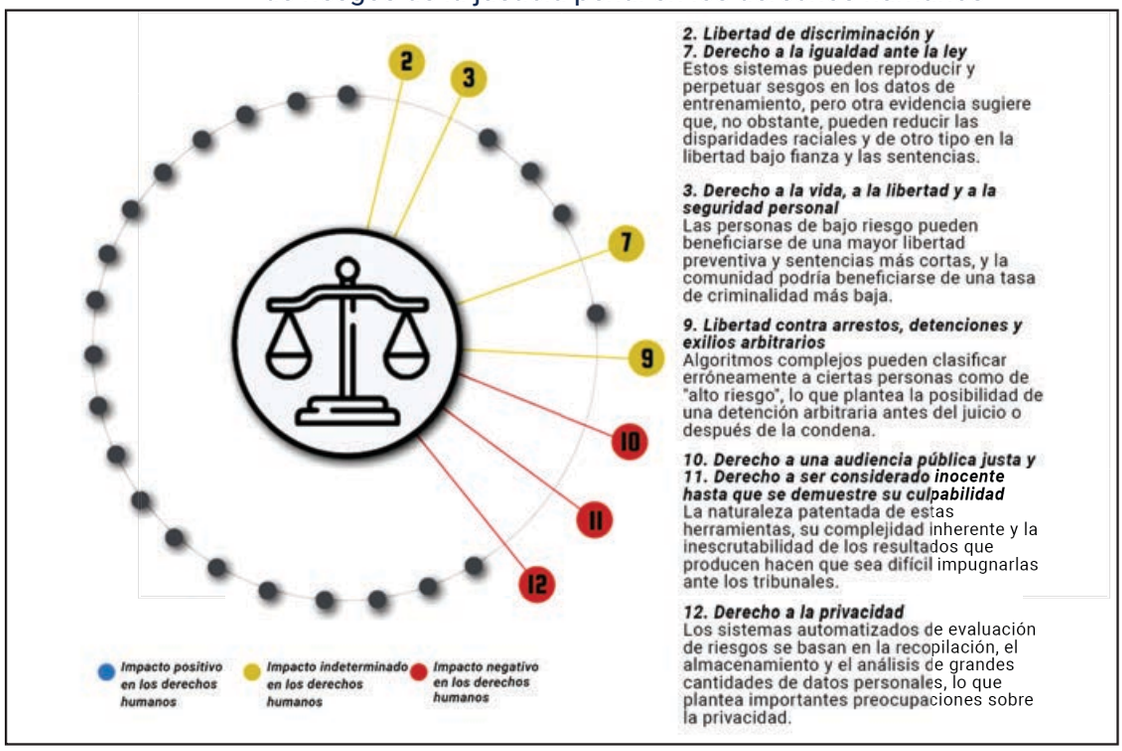
88 Comité de expertos en intermediarios de Internet (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Council of Europe Study, DGI/2017/12, disponible en: https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

89 Nir S. M. (2020). How 2 Lives Collided in Central Park, Rattling the Nation, disponible en: https://www.nytimes.com/2020/06/14/nyregion/central-park-amy-cooper-christian-racism.html

90 Heaven W. D. (2020). Predictive policing algorithms are racist. They need to be dismantled, disponible en: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

91 Disponible en: https://www.scc-csc.ca/case-dossier/cb/37233-eng.pdf

Figura 11. Impacto de las herramientas de evaluación de riesgos de la justicia penal en los derechos humanos



Fuente: Raso F., Hilligoss H., Krishnamurthy V., Bavitz C., Kim L. (2018). Artificial Intelligence & Human Rights: Opportunities & Risks, disponible en:

Resolución de conflictos

Los sistemas de IA se pueden utilizar para pronosticar cómo se decidirá un caso, proporcionando así a los demandantes una mejor comprensión de sus opciones o generando una propuesta de acuerdo. En este enfoque, la predicción de decisiones judiciales podría facilitar el acceso a la justicia. Dichos sistemas pueden integrarse en plataformas judiciales en línea donde las personas exploran sus alternativas legales o ingresan e intercambian información relacionada con el caso. El sistema de IA ayudaría a los litigantes a tomar mejores decisiones sobre la presentación de demandas y a los tribunales a acelerar la toma de decisiones complementando o sustituyendo las conclusiones de los jueces.[92]

Muchas plataformas de resolución de disputas en línea (ODR) no utilizan IA, sino que sirven como plataforma para la coordinación y simplificación del trabajo de los litigantes. Sin embargo, las plataformas ODR como Rechtwijzer, utilizadas en los Países Bajos[93], MyLaw BC, Canadá[94], y el ODR utilizado por el Tribunal de Resolución Civil de Columbia Británica (CRT), Canadá[95], utilizan sistemas de IA para determinar qué partes pueden utilizar la plataforma para resolver una disputa, así como para automatizar la toma de decisiones y la resolución o recomendación de resultados.

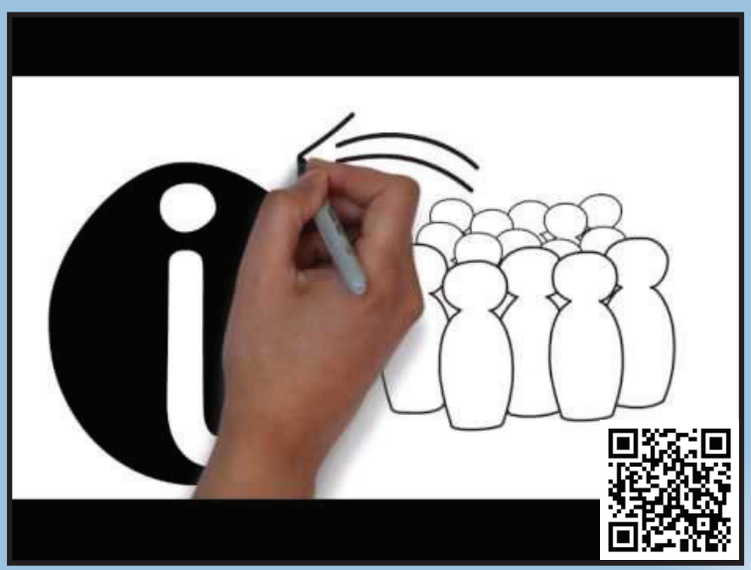
Por ejemplo, el procedimiento de resolución de disputas de CRT de Columbia Británica comienza con el Explorador de soluciones, un sistema experto en IA, que emplea una estructura de preguntas y respuestas para proporcionar

92 Wu J. (2019). AI Goes to Court: The Growing Landscape of AI for Access to Justice, disponible en: https://medium.com/legal-design-and-innovation/ai-goes-to-court-the-growing-landscape-of-ai-for-access-to-justice-3f58aca4306f
93 Véase: https://rechtwijzer.nl/
94 Véase: https://family.legalaid.bc.ca/retiring-mylawbc
95 Véase: https://civilresolutionbc.ca/

a las personas usuarias información legal individualizada y en un lenguaje simple y recursos de autoayuda gratuitos para resolver su problema sin la necesidad de presentar una demanda de CRT. Abogados de toda Columbia Británica contribuyeron a producir contenido legal para el Explorador de soluciones. Los ingenieros del conocimiento visitaron a los abogados y los entrevistaron sobre los problemas más frecuentes observados en sus áreas de práctica, así como los hechos jurídicos que creen que el público debe conocer. Luego, el equipo de CRT organizó estos datos en extensos mapas mentales, asegurándose de que el lenguaje y el contenido sean claros y sean comprensibles por lectores de sexto grado escolar.[96]

Actividad: El ejemplo del Explorador de soluciones CRT de Columbia Británica

Los participantes de la capacitación ven el video a continuación y debaten si se pueden encontrar soluciones similares que utilicen sistemas expertos en IA en sus jurisdicciones.



Fuente: https://youtu.be/ueVUETHy8gc

Estudio de caso: bot jurado

Cada año, el Tribunal Superior del Condado de Los Ángeles se ocupa de alrededor de 1,2 millones de nuevas multas de tráfico. Hace varios años, las personas tuvieron que esperar hasta 2 horas y media para ver a un empleado por su problema de tráfico debido a una crisis financiera estatal que resultó en cierres de juzgados y reducción de personal.[97] Ahora, un asistente en línea del Tribunal Superior de Los Ángeles ayuda a las personas con sus multas de tráfico. El bot jurado utiliza servicios de traducción de ML y comprensión del lenguaje natural. Asiste a más de 5.000 personas ciudadanas cada semana y habla cinco idiomas.

Fuente: The Superior Court of California, County of Los Angeles, disponible en: https://ww2.lacourt.org/traffic/ui/trafficOS.aspx?s=1&language=2

96 Salter S. (2018). What is the Solution Explorer?, disponible en: https://www.cbabc.org/BarTalk/Articles/2018/April/Features/What-is-the-Solution-Explorer

97 SRLN (2023). Noticias: Gina - Avatar de tráfico en línea de Los Ángeles cambia radicalmente la experiencia del cliente (Los Ángeles 2016), disponible en: https://www.srln.org/node/1186/gina-las-online-traffic-avatar-radically-changes-customer-experience-news-2016

En Australia, el estado de Victoria está probando plataformas ODR a través de su piloto VCAT para demandas menores.[98] Estos pilotos utilizan plataformas como Modria, Modron y Matterhorn de Court Innovations. No está claro en qué medida se incluye la IA en estos sistemas, pero parecen ser principalmente plataformas para registrar hechos y preferencias, la interacción entre las partes y la redacción/firma de acuerdos (sin que ningún algoritmo o herramienta de IA decida o elabore una estrategia para las partes). Si los pilotos tienen éxito y se convierten en iniciativas en curso, las iteraciones futuras pueden incluir recomendaciones adicionales impulsadas por IA o ayudas para la toma de decisiones.[99]

IA generativa

El campo de la IA generativa está experimentando actualmente una era de progreso sin precedentes. Estos algoritmos de aprendizaje automático se han diseñado para crear contenido nuevo, como audio, código, imágenes, texto, simulaciones y vídeos. Recientemente, se han desarrollado chatbots como ChatGPT, Bard y Copilot que utilizan modelos de lenguaje de gran tamaño (LLM) para realizar diversas funciones, como la recopilación de investigaciones, la compilación de archivos de casos legales, la automatización de tareas administrativas repetitivas y la búsqueda en línea. Esta tecnología innovadora tiene el potencial de aumentar significativamente la eficiencia y la productividad al simplificar procesos y decisiones específicos, como agilizar el procesamiento de notas o ayudar a los educadores a enseñar habilidades de pensamiento crítico.[100]

Punto de debate: los participantes de la capacitación ven el video y debaten cómo la IA generativa ha influido en sus vidas. ¿Han intentado usarlo en los procesos de toma de decisiones? ¿Cuáles son las oportunidades y desafíos clave relacionados con la IA generativa?

Fuente: https://www.youtube.com/watch?v=hflUstzHs9A

98 Legaltech News (2020). A Future ODR Roadmap for Courts Post-COVID-19, disponible en: https://www.law.com/legaltechnews/2020/06/23/a-future-odr-roadmap-for-courts-post-covid-19/
99 Ibid.
100 Routley N. (2023). What is generative AI? An AI explains, disponible en: https://www.weforum.org/agenda/2023/02/generative-ai-explain-algorithms-work/.

Los sistemas de IA generativa pueden generar texto, incluidos argumentos legales o investigación, prediciendo el texto apropiado para seguir una entrada dada utilizando patrones aprendidos de conjuntos de datos extensos. Esto hace que la IA generativa sea una herramienta potente en varios campos, incluida la profesión legal. Mientras que algunas herramientas generativas de IA operan dentro de un universo cerrado de información, otras son abiertas y tienen un acceso más amplio a los datos, como a través de complementos web o conexiones a Internet.[101]

Muchos gobiernos de todo el mundo han comenzado a restringir el uso de modelos de lenguaje de gran tamaño (LLM)[102]. El borrador de la Ley de IA de la UE también contiene reglas para la IA de propósito general, o sistemas de IA que pueden implementarse para una variedad de tareas con varios niveles de riesgo. Tecnologías similares incluyen ChatGPT y otros sistemas de IA generativa LLM. En otro ejemplo, debido a problemas de protección de datos y privacidad, el regulador de protección de datos italiano emitió una prohibición temporal de ChatGPT.[103]

Los LLM como ChatGPT recopilan grandes cantidades de datos de Internet, incluida la información personal. El gobierno canadiense ha adoptado un enfoque proactivo para regular el uso de la IA generativa al publicar un borrador de un código de prácticas, que ahora está abierto a comentarios públicos. El código se promulgará como ley como parte de la Ley de Inteligencia Artificial y Datos del país.[104]

Mientras tanto, el G7 ha lanzado el Proceso de IA de Hiroshima para coordinar los debates sobre los riesgos asociados con la IA generativa.[105] En julio de 2023, el presidente de los Estados Unidos, Joe Biden, anunció compromisos voluntarios de las grandes empresas de IA para priorizar la seguridad y la confianza.[106] El 13 de julio de 2023, China implementó medidas temporales para regular la industria de IA generativa. Las nuevas reglas exigen que los proveedores de servicios se sometan a evaluaciones de seguridad y algoritmos de archivo para su revisión.[107] Además, la Autoridad de Salud Municipal de Beijing ha propuesto 41 nuevas normas que prohíben estrictamente el uso de IA en diversas actividades de atención médica en línea, incluida la generación automática de recetas médicas.[108]

101 Perkins Coie (2023). Use of Generative AI in Litigation Requires Care and Oversight, disponible en: https://www.perkinscoie.com/en/news-insights/use-of-generative-ai-in-litigation-requires-care-and-oversight.html.

102 Definición de LLM por Tech Target: "Un modelo de lenguaje de gran tamaño (LLM) es un tipo de algoritmo de inteligencia artificial (IA) que utiliza técnicas de aprendizaje profundo y conjuntos de datos muy grandes para comprender, resumir, generar y predecir nuevo contenido. El término IA generativa también está estrechamente relacionado con los LLM, que son, de hecho, un tipo de IA generativa que se ha diseñado específicamente para ayudar a generar contenido basado en texto", consulte: https://www.techtarget.com/whatis/definition/large-language-model-LLM

103 McCallum S. (2023). ChatGPT banned in Italy over privacy concerns, disponible en: https://www.bbc.com/news/technology-65139406

104 Canadian Guardrails for Generative AI – Code of Practice (2023), disponible en: https://ised-isde.canada.ca/site/ised/en/consultation-development-canadian-code-practice-generative-artificial-intelligence-systems/canadian-guardrails-generative-ai-code-practic

105 La Casa Blanca (2023). G7 Hiroshima Leaders' Communiqué, disponible en: https://www.whitehouse.gov/briefing-room/statements-releases/2023/05/20/g7-hiroshima-leaders-communique/

106 Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI, disponible en: https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/

107 Reuters (2023). China says generative AI rules to apply only to products for the public, disponible en: https://www.reuters.com/technology/china-issues-temporary-rules-generative-ai-services-2023-07-13/

108 Beijing to limit use of generative AI in online healthcare activities, including medical diagnosis, amid growing interest in ChatGPT-like services, disponible en: https://www.scmp.com/tech/policy/article/3231828/beijing-limit-use-generative-ai-online-healthcare-activities-including-medical-diagnosis-amid

En otras noticias, la Comisión Federal de Comercio de EE. UU. (FTC) ha iniciado una investigación sobre OpenAI por acusaciones de violaciones de la ley de protección al consumidor. La Demanda de Investigación Civil de la FTC ha expresado su preocupación de que ChatGPT, un modelo de lenguaje desarrollado por OpenAI, pueda producir declaraciones falsas o despectivas sobre personas reales. La agencia también ha solicitado información tras una violación de la privacidad de los datos en la que se expusieron datos privados de las personas usuarias en los resultados de ChatGPT.[109]

El ejemplo de ChatGPT

ChatGPT (Generative Pre-trained Transformer) es un chatbot que aprovecha el procesamiento avanzado del lenguaje natural (PNL) y el aprendizaje de refuerzo para participar en diálogos realistas con las personas. ChatGPT puede generar artículos, cuentos, poesía e incluso código informático. También puede responder a preguntas, participar en debates y, en ciertos casos, proporcionar respuestas extensas a preguntas e indagaciones extremadamente precisas. ChatGPT se lanzó en noviembre de 2022 y adquirió más de un millón de usuarios en una semana.[110]

El poder judicial no ha sido inmune a las controversias relacionadas con el uso de la IA generativa. Por ejemplo, en enero de 2023 se presentó una controversia en Colombia después de que un juez revelara que utilizó ChatGPT para ayudarlo a determinar si el seguro de un niño autista debería cubrir todos los gastos relacionados con su tratamiento médico.[111] Diez días después de este controvertido fallo, todavía en Colombia, un magistrado emitió una orden judicial utilizando ChatGPT para ayudarlo a decidir cómo llevar a cabo un juicio en el metaverso. Además, a fines de marzo de 2023, un juez en Perú y un magistrado en México afirmaron haber utilizado el ChatGPT de OpenAI para motivar una decisión de segunda instancia y para ilustrar sus argumentos en una audiencia judicial.[112]

Siguiendo el caso Mata vs. Avianca Airlines, Inc.[113], donde un abogado presentó citaciones falsificadas y casos creados por ChatGPT a un tribunal de los Estados Unidos, las pautas de uso responsable se han vuelto aún más vitales. El juez federal Brantley Starr (Distrito Norte de Texas) implementó una nueva norma que exige una certificación más explícita y precisa. Esta certificación garantiza que cualquier texto generado por IA generativa se someterá a una verificación de precisión humana utilizando fuentes legales autorizadas antes de presentarse al Tribunal.[114] Su pedido requería lo siguiente:

109 Reuters (2023). US FTC opens investigation into OpenAI over misleading statements, disponible en: https://www.reuters.com/technology/us-ftc-opens-investigation-into-openai-washington-post-2023-07-13/
110 https://chat.openai.com/
111 Gutiérrez J. D. (2023). ChatGPT in Colombian Courts: Why we need to have a conversation about the digital literacy of the judiciary, disponible en: https://verfassungsblog.de/colombian-chatgpt/
112 Gutiérrez J. D. (2023). Los jueces y magistrados en Perú y México tienen fiebre de ChatGPT, disponible en: https://techpolicy.press/judges-and-magistrates-in-peru-and-mexico-have-chatgpt-fever/
113 Mata vs. Avianca, Inc., 1:22 -cv-01461, disponible en: https://www.courtlistener.com/docket/63107798/mata-v-avianca-inc/
114 Hunton Andrews Kurth (2023). Will Mandatory Generative AI Use Certifications Become The Norm In Legal Filings?, disponible en: https://www.huntonak.com/en/insights/will-mandatory-generative-ai-use-cerfifications-become-the-norm-in-legal-filings.html. Consulte también: https://law.mit.edu/ai

"Todos los abogados y litigantes en causa propia que comparezcan ante el Tribunal deben, junto con su notificación de comparecencia, presentar en el expediente un certificado que acredite que ningún fragmento de cualquier presentación será redactado por inteligencia artificial generativa (como ChatGPT, Harvey.AI o Google Bard) o que cualquier lenguaje redactado por inteligencia artificial generativa será verificado por un ser humano para revisar su precisión, utilizando reportes impresos o bases de datos legales tradicionales. Estas plataformas son increíblemente poderosas y tienen muchos usos en la ley: formar divorcios, solicitudes de descubrimiento, errores sugeridos en documentos, preguntas anticipadas en debates orales. Pero la exposición legal no es uno de ellos. He aquí la razón. Estas plataformas en sus estados actuales son propensas a alucinaciones y prejuicios. En las alucinaciones, inventan cosas, incluso citas. Otro problema es la fiabilidad o el sesgo. Mientras que los abogados juran dejar de lado sus prejuicios personales, prejuicios y creencias para defender fielmente la ley y representar a sus clientes, la inteligencia artificial generativa es el producto de una programación ideada por humanos que no tuvieron que hacer tal juramento. Como tales, estos sistemas no tienen lealtad a ningún cliente, al Estado de derecho o a las leyes y la Constitución de los Estados Unidos (o, como se mencionó anteriormente, a la verdad). Sin ningún sentido del deber, el honor o la justicia, dichos programas actúan de acuerdo con el código informático en lugar de la convicción, basados en la programación en lugar de los principios. Cualquier parte que crea que una plataforma tiene la precisión y confiabilidad requeridas para la información legal puede solicitar una licencia y explicar por qué. En consecuencia, el Tribunal anulará cualquier radicación de una parte que no presente un certificado en el expediente que acredite que ha leído los requisitos específicos del juez del Tribunal y entiende que será responsable en virtud de la Regla 11 por el contenido de cualquier radicación que firme y presente al Tribunal, independientemente de si la inteligencia artificial generativa redactó alguna parte de esa presentación".

Fuente: https://www.txnd.uscourts.gov/judge/judge-brantley-starr

Estos son tres riesgos principales de la IA generativa con respecto al poder judicial:

- Propósito/cambios imprevistos. Un sistema de IA diseñado e implementado para el propósito "A" no debe usarse a ciegas para alguna función alternativa. Por ejemplo, una herramienta de PNL principalmente para la traducción de órdenes judiciales no debe usarse arbitrariamente para ayudar también a las consultas de casos o ayudar a los jueces en la toma de decisiones sin revelar su uso para tales fines adicionales. En algunos casos, los propósitos adicionales pueden ser válidos, en otros no. Incluso cuando las funciones adicionales puedan considerarse legales y válidas, puede ser necesario entrenar el algoritmo base sobre datos relevantes adicionales para garantizar la precisión y la fiabilidad. Básicamente, una expansión ciega de los cambios imprevistos generalmente exacerba los riesgos potenciales de un sistema de IA de propósito general y debe ser disuadido o al menos regulado.

- Alucinaciones y desinformación. Es importante tener en cuenta que los modelos de IA generativa se entrenan con grandes cantidades de datos, lo que resulta en respuestas altamente realistas y relevantes. Sin embargo, vale la pena señalar que las herramientas que utilizan dichos modelos pueden producir resultados que son plausibles pero no del todo precisos debido a la naturaleza de su diseño, que tiene como objetivo generar resultados que se parezcan mucho pero que no sean idénticos a la información de origen. La IA de propósito general, especialmente los LLM, han demostrado cada vez más el potencial de "alucinar", es decir, dar resultados inexactos de una manera convincente similar a la humana, haciéndolos creíbles y aumentando el riesgo de su aceptación como precisos (una forma de sesgo de automatización). Esto es particularmente peligroso en el sistema judicial: hemos tenido diferentes instancias en los últimos meses de jueces que confían en ChatGPT para dar su opinión sobre la jurisprudencia existente con respecto a la cuestión legal. Esto se informó en Colombia en un caso de seguros, e incluso en la India (juez del Tribunal Superior de Punjab y Haryana). La salida alucinada puede resultar extremadamente problemática, especialmente para la resolución.
- Preocupaciones sobre Propiedad intelectual. Una vez más, los LLM deben considerarse dadas las preocupaciones sobre los derechos de propiedad intelectual tradicionales de los creadores de obras originales.

Actividad: Los participantes en la capacitación leen el texto a continuación sobre las implicaciones de los derechos de autor del uso de la IA generativa y debate si las doctrinas de "uso legítimo" o "excepciones permisibles de derechos de autor" podrían aplicarse en el contexto de la IA generativa.

Con el auge de la IA generativa, las demandas parecen estar convirtiéndose en algo cotidiano. En noviembre de 2022, Microsoft, GitHub y OpenAI se enfrentaron a una demanda colectiva señalando que el sistema Copilot propiedad de GitHub, que fue entrenado en miles de millones de líneas de código público, viola la ley de derechos de autor al emitir fragmentos de código con licencia sin atribución.[115] A cambio, las empresas argumentaron ante un tribunal federal en San Francisco que la demanda actual sobre el uso de código abierto para entrenar sus sistemas de IA no tiene sustento. Las empresas afirmaron que la denuncia carece de especificidad en sus alegaciones. Además, argumentaron que el sistema Copilot de GitHub, que proporciona sugerencias de código a los programadores, utiliza el código fuente de una manera consistente con los principios de uso legítimo.[116]

También hay un caso judicial contra Midjourney y Stability AI, las empresas responsables de las herramientas de arte de IA ampliamente utilizadas. El caso

115 Vincent J. (2022). The lawsuit that could rewrite the rules of AI copyright, disponible en: https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data

116 IT world Canada (2023). Microsoft, GitHub, and OpenAI ask court to dismiss AI copyright lawsuit, disponible en: https://www.itworldcanada.com/post/microsoft-github-and-openai-ask-court-to-dismiss-ai-copyright-lawsuit

afirma que estas empresas violaron los derechos de millones de artistas al utilizar imágenes web para entrenar sus herramientas.[117]

Además, Getty Images presentó una demanda contra Stability AI por supuestamente utilizar millones de imágenes de su sitio sin autorización para entrenar a Stable Diffusion, una IA capaz de generar arte.[118]

La principal preocupación con la IA generativa es su inclinación a replicar imágenes, texto y otros tipos de contenido, incluidos los que tienen derechos de autor, a partir de sus datos de entrenamiento. Este problema se destacó en un incidente reciente en el que se descubrió que una herramienta de IA utilizada por CNET para escribir artículos explicativos había plagiado artículos escritos por humanos, que probablemente formaban parte de su conjunto de datos de entrenamiento.[119] Además, un estudio académico de diciembre reveló que los modelos de IA capaces de generar imágenes, como DALL-E 2 y Stable Diffusion, pueden replicar ciertos elementos de imágenes a partir de sus datos de entrenamiento.[120]

Ciertas plataformas que alojan imágenes han prohibido el uso de contenido generado por IA debido a posibles repercusiones legales. Los profesionales judiciales también han advertido que el uso de herramientas generativas de IA puede exponer a las empresas a riesgos si integran inadvertidamente contenido protegido por derechos de autor producido por estas herramientas en sus productos para la venta.

Empresas como Stability AI y OpenAI, los creadores de ChatGPT, han argumentado que están protegidos por la doctrina del "uso legítimo", incluso si sus sistemas se entrenaron con contenido con licencia. Este principio legal, reconocido en los Estados Unidos, permite el uso limitado de material protegido por derechos de autor sin obtener el permiso del propietario de los derechos. Los defensores del uso legítimo a menudo citan el ejemplo de Authors Guild vs. Google, donde el Tribunal de Apelaciones de los Estados Unidos para el Segundo Circuito en Nueva York determinó que el escaneo manual de Google de millones de libros con derechos de autor para desarrollar su plataforma de búsqueda de libros era un uso legítimo, incluso sin una licencia. Sin embargo, el concepto de uso legítimo se debate y modifica con frecuencia, y sigue sin probarse en gran medida en el ámbito de la IA generativa.[121]

Si las obras producidas por la IA pueden protegerse por la defensa del "uso legítimo" depende de si se consideran transformadoras. Esto significa que las obras deben utilizar materiales con derechos de autor de una manera que difiera significativamente de los originales. Casos legales anteriores, como la decisión de Google vs. Oracle de la Corte Suprema de los Estados Unidos en 2021, indican que la creación de nuevas obras a partir de los datos recopilados puede ser transformadora. El tribunal determinó que el uso por parte de Google de partes del código Java SE para desarrollar su sistema operativo Android se consideraba uso legítimo.[122]

Fuente: Tech Crunch (2023). The current legal cases against generative AI are just the beginning, disponible en: https://techcrunch.com/2023/01/27/the-current-legal-cases-against-generative-ai-are-just-the-beginning/

117 Vincent J. (2023). AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit, disponible en: https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart

118 Brittain B. (2023). Getty Images lawsuit says Stability AI misused photos to train AI, disponible en: https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/

119 Futurism (2023). CNET's AI Journalist Appears to Have Committed Extensive Plagiarism, disponible en: https://futurism.com/cnet-ai-plagiarism

120 Somepalli G., Singla V., Goldblum M., Geiping J., Goldstein J. (2022). Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models, Universidad de Maryland, disponible en: https://arxiv.org/pdf/2212.03860.pdf

121 Authors Guild vs. Google, Inc., No. 13-4829 (2d Cir. 2015), disponible en: https://law.justia.com/cases/federal/appellate-courts/ca2/13-4829/13-4829-2015-10-16.html

122 Setty R. (2023). First AI Art Generator Lawsuits Threaten Future of Emerging Tech, disponible en: https://news.bloomberglaw.com/ip-law/first-ai-art-generator-lawsuits-threaten-future-of-emerging-tech

Reconocimiento y análisis de idiomas

El uso de técnicas de IA puede reducir la necesidad de traducción humana. Estas herramientas pueden identificar rápidamente los documentos que contienen texto en idiomas extranjeros y proporcionar una lista de los idiomas que contienen, lo que permite una planificación más exhaustiva. Varias tecnologías de IA también pueden traducir texto de un idioma a otro.

Procesamiento del lenguaje natural (PLN)

El PLN es una técnica de aprendizaje automático que analiza grandes cantidades de texto humano o datos del habla (transcritos o acústicos) en busca de propiedades específicas, como significado, contenido, intención, actitud y contexto.[123]

El análisis del lenguaje se ha utilizado en el dominio judicial y la criminología durante mucho tiempo. Por ejemplo, la clasificación de textos se ha utilizado en lingüística forense. Mientras que en el pasado el análisis se hacía manualmente, hoy en día los métodos de ML se utilizan para identificar el género, la edad, los rasgos de personalidad e incluso la identidad de un autor, o para la transcripción en vivo.[124] Por ejemplo, la PNL puede ayudar a los operadores judiciales a identificar y vincular referencias a la misma persona u organización a lo largo de un conjunto de contratos legales. También se puede utilizar para analizar una colección de casos judiciales para identificar temas o cuestiones legales recurrentes, o para extraer los nombres de las partes involucradas, las fechas y los lugares mencionados en una opinión judicial. Además, los sistemas de PNL se pueden utilizar para eliminar automáticamente la información confidencial de los documentos judiciales, como los números de la seguridad social y las direcciones personales, a fin de proteger la privacidad de las personas.

Cabe señalar que los modelos de PNL siguen estando expuestos a errores, y los errores en la traducción pueden tener graves consecuencias para los derechos fundamentales de las personas cuando estos modelos se implementan en operaciones judiciales.

123 Firth-Butterfield K., Silverman K. (2022). Artificial Intelligence and the Courts: Materials for Judges. Artificial Intelligence – Foundational Issues and Glossary, American Association for the Advancement of Science, disponible en https://doi.org/10.1126/aaas.adf0782

124 Medvedeva M., Vols M., Wieling M. (2020). Using machine learning to predict decisions of the European Court of Human Rights, Artif Intell Law, 28, 237–266, disponible en: https://link.springer.com/article/10.1007/s10506-019-09255-y

Caso de estudio

SUVAS de la India

El Vidhik Anuvaad Software (SUVAS), un programa de inteligencia artificial que traduce decisiones y órdenes a nueve idiomas locales diferentes, fue presentado por el Tribunal Supremo en noviembre de 2019. SUVAS tenía como objetivo facilitar a las personas que no hablan inglés la obtención de sentencias y órdenes y ayudarles a comprender mejor los procedimientos judiciales.

Fuente: Press Trust of India, Software developed to translate SC judgments in 9 vernacular languages: Law Minister RS Prasad, disponible en: https://www.business-standard.com/article/pti-stories/software-developed-to-translate-sc-judgments-in-9-vernacular-languages-law-minister-rs-prasad-119121200851_1.html.

En febrero de 2023, Technology Enabled Resolution (TERES), una startup tecnológica con sede en Bangalore, India, comenzó a usar IA para iniciar la transcripción en vivo de las audiencias de la Corte Suprema.

Fuente: Mint (2023). Bangalore techies bring AI to Supreme Court for the first time, disponible en: https://www.livemint.com/news/india/supreme-court-uses-ai-based-transcript-for-the-first-time-here-s-how-it-works-11677403522929.html.

India ha tenido éxito en la creación de sus propios modelos de roles retóricos y NER entrenados en el texto legal indio. El modelo Ner específicamente tiene una precisión del 91 %.

Fuente: https://github.com/OpenNyAI/Opennyai

Gestión de casos y archivos digitales

La IA también podría facilitar la gestión de archivos digitales, lo que a su vez haría que los operadores judiciales fueran más efectivos al permitirles centrarse en asuntos más importantes.

Intelligent Trial 1.0, una IA para la gestión de tribunales en China

Por ejemplo, el Tribunal Superior de Hebei en China ha desarrollado Intelligent Trial 1.0, una IA inteligente de gestión judicial. Escanea y digitaliza automáticamente los expedientes; clasifica los documentos en expedientes electrónicos; empareja a las partes con las partes de los casos existentes; identifica las leyes, los casos y los documentos jurídicos pertinentes que deben tenerse en cuenta; genera todos los documentos procesales necesarios para el tribunal, como notificaciones y sellos; y distribuye los casos a los jueces para que puedan ponerse en marcha. La tecnología coordina numerosas tareas de IA en un flujo de trabajo que puede minimizar las cargas del personal judicial y los jueces.

Herramienta para la anonimización de documentos legales, Argentina

Para acelerar el proceso judicial y reducir el margen de error, Cambá Cooperative, una cooperativa laboral de bancos de software, ha creado un sistema de IA escalable para anonimizar documentos legales en español. El sistema de IA tiene como objetivo anonimizar los datos personales de los documentos públicos, reducir el tiempo y los errores, y salvaguardar el derecho a la privacidad. El Juzgado de lo Penal nº 10 de Buenos Aires, Argentina, implementó esta herramienta de IA en sus sentencias.[125]

125 Véase: https://www.empatia.la/en/proyecto/ia2/; ver también: Selvood I., Uribe P. (2022). Open Justice is Moving Forward in the Americas, disponible en: https://www.opengovpartnership.org/stories/open-justice-is-moving-forward-in-the-americas/

En profundidad: IA como prueba en procedimientos judiciales

La naturaleza compleja de los algoritmos de ML y su naturaleza opaca plantean desafíos para usar los sistemas de IA como evidencia en los procedimientos legales. Los tribunales deben establecer un método confiable para verificar la precisión de los resultados de la IA, lo que puede implicar el testimonio de expertos o medios técnicos como marcas de agua incrustadas en imágenes. Decidir quién está calificado para testificar sobre la precisión de las aplicaciones de IA también es un tema fundamental, con opciones que van desde ingenieros de software e ingenieros de diseño hasta ingenieros de datos y directores generales de empresas.[126]

Los jueces tienen dificultades para determinar la precisión de las herramientas de diagnóstico impulsadas por IA. Si bien la IA de diagnóstico médico se puede comparar con los diagnósticos médicos, no está claro cómo se pueden evaluar científica o evidentemente los algoritmos diseñados para predecir el comportamiento futuro, como las herramientas de evaluación penal. Puede ser difícil determinar la causalidad con algoritmos predictivos en el contexto criminal, ya que también consideran los factores sociales que pueden influir en el comportamiento. La evaluación de la precisión, las tasas de error y la realización de pruebas y la revisión por pares son tareas fundamentales pero difíciles en este campo. Una vez que una persona ha sido encarcelada o sentenciada, se hace difícil predecir cómo su comportamiento futuro puede haber sido influenciado por su encarcelamiento. Los efectos del encarcelamiento, incluido el apoyo de los seres queridos en el exterior, pueden tener un impacto significativo en el comportamiento futuro de una persona, lo que hace extremadamente difícil medir con certeza la precisión de la predicción de ML.

Las partes en controversia también buscarán desafiar la relevancia y precisión del sistema de ML buscando acceso al algoritmo subyacente, los datos sobre los que fue entrenado, validado y probado, así como lo que ocurre y se pondera dentro de cualquier caja negra de aprendizaje automático. Por lo tanto, los tribunales podrían enfrentar desafíos decisorios en capas cada vez que se ofrezca evidencia generada por IA. Cuando se admitan resultados de IA, la contraparte intentará interrogar a los ingenieros de software responsables de su diseño. Además, debido a que cada aplicación de IA es diferente, podrá:

- Tener diferentes propósitos de salida;
- Depender de diferentes algoritmos;
- Utilizar diferentes metodologías de aprendizaje automático;
- Entrenar, probar y validar utilizando diferentes datos.

Estas cuestiones generalmente no están sujetas a resolución mediante la aplicación de precedentes jurisprudenciales de la misma manera, por ejemplo, que el análisis de ADN ahora es generalmente aceptado en los tribunales. Se debe esperar la decisión sobre cada solicitud y en cada contexto para el cual la solicitud se ofrece como prueba.

126 Baker J. E., Hobart L. N., Mittelstead M. G. (2021). AI for Judges. A Framework. Centro de Seguridad y Tecnología Emergente, disponible en: https://www.armfor.uscourts.gov/ConfHandout/2022ConfHandout/Baker2021DecCenterForSecurityAndEmergingTechnology1.pdf

Oportunidades: la IA y el poder judicial en África

Los rápidos avances en las tecnologías de IA y PNL presentan nuevas posibilidades para modernizar el sector de la justicia en África. Por ejemplo, empresas como Juta[127] en Sudáfrica están aprovechando estas innovaciones para desarrollar soluciones de vanguardia que ayuden a los bufetes de abogados y otras organizaciones legales a realizar investigaciones legales exhaustivas y descubrir recursos valiosos para sus casos.[128] Al capitalizar el vasto archivo de documentos legales de Juta y utilizar técnicas analíticas avanzadas, los sistemas judiciales africanos pueden mejorar su eficiencia y eficacia.

Datos del caso

Un área potencial en la que la tecnología de IA podría incorporarse a los sistemas judiciales africanos es a través de la digitalización de los datos de los casos judiciales. Al capturar información detallada sobre diversos aspectos del proceso legal, incluidas sentencias, fallos, decisiones, antecedentes de casos, partes involucradas, etc., permitiría a los algoritmos de aprendizaje profundo identificar patrones e ideas a partir de estos datos. Organizar y almacenar adecuadamente los datos de casos recopilados en grandes bases de datos ayudará a establecer una base que permita a los africanos aprovechar su valor para una multitud de aplicaciones y funciones procesables. La precisión del sistema de registro también debe ser auditable y se debe dar prioridad a la prueba física sobre los registros digitales. Esta no es una tarea fácil y requerirá una gran coordinación en el sistema judicial. Los análisis tan sencillos como el seguimiento del progreso de los diferentes tribunales frente a los predecesores de cada juez en años anteriores podrían determinar a qué juez asignar ciertos tipos de juicios en función de la productividad en períodos objetivo promedio mediante estudios de correlación de alta tasa de éxito. Otra arista sería examinar declaraciones legales desde dentro de un caso que es de código abierto o conjuntos de datos generados por el gobierno.

Manejo del descubrimiento y la recuperación de información

Para mejorar la eficiencia de la fase de descubrimiento en los procedimientos legales y facilitar un intercambio más efectivo de la documentación relevante entre las partes interesadas, la implementación de archivos digitales es fundamental.[129] Al establecer una plataforma en línea para almacenar archivos y pruebas esenciales, los sistemas judiciales pueden aprovechar los mecanismos de búsqueda de vanguardia para localizar información importante de manera rápida y precisa. Este enfoque no solo agiliza la gestión de la información, sino que también permite a los abogados estructurar argumentos más sólidos respaldados por hechos confiables derivados de fuentes accesibles e interconectadas.

127 Juta and Company es un proveedor líder de contenido legal, regulatorio, comercial y académico de calidad en toda África; consulte: https://juta.co.za

128 Jutastat Evolve es una solución de investigación analítica cognitiva para un descubrimiento rápido y preciso, información y análisis de datos; consulte https://jutastatevolve.co.za/

129 Kufakwababa C. Z. (2021). Herramientas de inteligencia artificial en la automatización del trabajo jurídico: el uso y la percepción de las herramientas para el descubrimiento de documentos y los procesos de clasificación de privilegios en los bufetes de abogados del sur de África, Tesis doctoral, Stellenbosch: Stellenbosch University.

Uso de procedimientos judiciales multimodales

La modernización de los entornos de las salas de audiencias a través de la reunión de medios multipropósito podría beneficiar enormemente las operaciones judiciales en toda África. La integración de una gama de entradas sensoriales, incluidas las grabaciones de audio y vídeo, ofrece varias ventajas. Los avances tecnológicos en la visión artificial y la escucha de máquinas pueden mejorar sustancialmente la forma en que los transcriptores convierten las palabras habladas en texto, disminuyendo el error humano al tiempo que aumentan la velocidad y la precisión. Estas transcripciones digitales se convierten en herramientas para el análisis de investigación posterior al juicio y, cuando se combinan con capacidades de modelado predictivo, allanan el camino para un apoyo a la toma de decisiones más sofisticado durante las audiencias activas. Además, la indexación de los activos multimedia para favorecer el acceso facilita tanto la referencia judicial como el escrutinio público, lo que contribuye a una mayor confiabilidad dentro del marco legal. Los responsables de la toma de decisiones podrían considerar la implementación de proyectos piloto utilizando métodos inteligentes de mantenimiento de registros y observando resultados prometedores antes de una adopción más amplia de formatos multimedia abiertos. Luego, los cambios sistémicos pueden enfocarse a prioridades nacionales específicas.

Mejora de las herramientas lingüísticas para el poder judicial

El uso de tecnologías avanzadas de procesamiento del lenguaje natural, como la traducción automática[130] y la clasificación de documentos, ofrece una excelente oportunidad para que los tribunales africanos aborden las barreras lingüísticas. La falta de apoyo en el idioma local dificulta la participación pública y la difusión de información vital sobre los procedimientos judiciales. La adopción de soluciones modernas de IA para las traducciones garantiza un acceso equitativo a los recursos legales en diversas poblaciones con diferentes idiomas nativos. Mientras tanto, la clasificación del contenido en el idioma local faculta al sistema de justicia para aceptar y analizar presentaciones multiculturales, reduciendo las divisiones geográficas entre intérpretes, litigantes y personal judicial. Por ejemplo, los informes académicos publicados por asociaciones profesionales enfatizan que eliminar la discriminación lingüística y promover la paridad dentro del ámbito legal podría aliviar problemas similares relacionados con la jurisprudencia en

130 Adelani D., Alabi J., Fan A., Kreutzer J., Shen X., Reid M., Ruiter D., Klakow D., Nabende P., Chang E., Gwadabe T. (2022). A Few Thousand Translations Go a Long Way! Aprovechamiento de modelos preentrenados para la traducción de noticias africanas, en actas de la Conferencia 2022 de la Sección Norteamericana de la Asociación de Lingüística Computacional: Tecnologías del lenguaje humano, 3053–3070.

África.[131] Dado el mayor interés dirigido hacia el desarrollo de léxicos regionales y técnicas inferenciales, más naciones pueden ponderar las presentaciones constituyentes personalizadas. Posteriormente, los gobiernos demostrarían un compromiso tangible con la integración constructiva de las zonas remotas.

Archivos legales abiertos

Los archivos abiertos que contienen colecciones completas de decisiones judiciales continentales sirven como recursos valiosos tanto para los profesionales del derecho como para los académicos. Aspectos como la facilidad de navegación amplifican la importancia de estas bases de datos para promover deliberaciones informadas. Si bien algunas naciones africanas han logrado avances significativos en la digitalización de sus fallos en los tribunales superiores, los tribunales inferiores siguen estando comparativamente subrepresentados. A pesar de tener organizaciones dedicadas a la tecnología legal, tal distribución desequilibrada merece atención. Por lo tanto, es necesario mejorar la infraestructura de tecnología de la información para las instituciones judiciales a fin de garantizar una cobertura uniforme de todos los tribunales, fomentando una accesibilidad equilibrada y la igualdad de oportunidades para el avance a través de la información de datos legales.

Conexión con la IA local

Las instituciones académicas africanas y las instalaciones de investigación privadas centradas en la Inteligencia Artificial (IA) deben buscarse por el poder judicial para reforzar las colaboraciones conjuntas que maximicen los beneficios derivados de estas asociaciones. El fomento de estas interacciones ayuda a navegar por marcos regulatorios internacionales complejos a través de conocimientos y experiencias compartidos. Además, la participación en prolíficas iniciativas locales de la comunidad de IA como Deep Learning Indaba, Data Science Africa, Masakhane Research Organisation, Data Science Network, que se compone de numerosos investigadores repartidos en varias naciones, podría mejorar en gran medida la conexión de los sistemas judiciales con mentes innovadoras dentro de la región. Por lo tanto, adoptar la colaboración en todo el continente posee un potencial transformador que abarca la competencia técnica dentro de los tribunales y la inclusión social en general.

131 Docrat Z., (2022). Una revisión de las cualificaciones lingüísticas y la formación de los Profesionales del Derecho y los Funcionarios judiciales: Un llamamiento a la igualdad lingüística en la abogacía sudafricana, Revista Internacional de Semiótica Jurídica-Revue internationale de Sémiotique juridique, 35(5), 1711–1731.

2. Casos de estudio sobre el despliegue de IA en el poder judicial

Esta sección ofrece una descripción general de casos seleccionados de despliegue de IA en el poder judicial en Brasil, Singapur, Argentina, Colombia, India, Reino Unido y Estados Unidos. Cabe señalar que esto no sirve como respaldo de estos casos de uso de la IA en determinados poderes judiciales nacionales, y que los operadores judiciales deben ser conscientes de todos los riesgos (sesgos, cajas negras, ciberseguridad e invasión de los derechos humanos) que podrían ocurrir con el uso de sistemas de IA en las operaciones judiciales.

VICTOR, Brasil

El Tribunal Supremo de Brasil (STF) utiliza el sistema VICTOR AI, que se desarrolló en colaboración con la Universidad de Brasilia (UnB). La tecnología de IA analiza el enorme volumen de apelaciones presentadas ante el Tribunal Superior y automatiza el proceso de examen mediante la identificación de casos con repercusión general, un requisito para la tramitación de una apelación ante el STF.

Solo en 2018, se presentaron más de cincuenta mil apelaciones ante este Tribunal, que tiene el potencial de decidir alrededor de ciento veinte mil casos anualmente. La primera etapa en el análisis de todos los recursos que llegan al STF es determinar si tienen repercusiones generales. Antes de VICTOR, este análisis era realizado por funcionarios judiciales sobre la base de los precedentes vinculantes de los jueces, y tomaba alrededor de cuarenta minutos por caso.

En cuanto a su diseño de software, VICTOR incorpora varias tecnologías de vanguardia y una gran base de datos de documentos judiciales. El conjunto de datos utilizado para capacitar a VICTOR contiene más de 100.000 demandas y casi tres millones de expedientes de casos extraídos durante un período de dos años (2017-2019).

Su problema inicial era lidiar con la realidad de que los documentos judiciales de todos los tribunales brasileños (Estado, Federal, Laboral, Militar, Justicia Electoral) llegan al STF en formatos variados, como volúmenes PDF no estructurados que contienen documentos no indexados.[132]

Sistema inteligente de transcripción judicial de Singapur

El Sistema inteligente de transcripción judicial (iCTS) se ha implementado en los tribunales de Singapur en asociación con el Instituto de Investigación Infocomm de A*STAR. El iCTS tiene el potencial de aumentar la eficiencia de la corte al transcribir las audiencias judiciales en tiempo real, eliminando

132 Salomao L. F., Braga R. (2020). El papel del Poder judicial en la realización de la Agenda 2030 de la ONU, disponible en: https://www.conjur.com.br/2021-jul-09/salomao-braga-judiciario-agenda-2030-onu. Véase también: https://portal.fgv.br/en/news/artificial-intelligence-judiciary-and-its-role-implementing-un-agenda-2030; https://sifocc.org/app/uploads/2020/06/Victor-Beauty-or-the-Beast.pdf

la necesidad de contratar a un transcriptor humano y permitiendo que los jueces y las partes revisen los testimonios orales en la corte de inmediato. Lo hace mediante el uso de redes neuronales entrenadas con modelos de lenguaje y términos específicos del dominio (como la terminología legal).[133]

Cabe señalar que los sistemas de reconocimiento de voz tienen una "reputación" de no funcionar bien cuando se exponen a ciertos acentos, lo que termina siendo discriminatorio en ciertas circunstancias. Los operadores judiciales deben ser conscientes de estas deficiencias.

Prometea, Argentina

El sistema Prometea utiliza enfoques de IA para generar opiniones judiciales automáticamente. En 2017, la Fiscalía de la Ciudad Autónoma de Buenos Aires, Argentina, comenzó a desarrollar Prometea. La herramienta ha permitido a la Fiscalía mejorar significativamente la eficiencia de sus procesos: una reducción de 90 minutos a un minuto (99 %) para la resolución de un proceso de licitación y de 167 días a 38 días (77 %) para la preparación del juicio.[134]

Prometea se distingue por tres características principales:

- Ofrece una interfaz intuitiva y fácil de usar que permite el reconocimiento del lenguaje natural y "hablar" con la máquina. En una sola pantalla, la persona usuaria tiene acceso a todos sus recursos relacionados con el trabajo.
- Funciona como un sistema experto multifuncional con la capacidad de automatizar el procesamiento de documentos y proporcionar soporte inteligente.
- Emplea enfoques supervisados de aprendizaje automático y agrupación, basados en el etiquetado manual y la capacitación en conjuntos de datos generados por máquinas.[135]

Las funcionalidades de Prometea se pueden dividir en cuatro categorías:

- Asistencia inteligente: Prometea ayuda a los responsables de la toma de decisiones y a las personas usuarias a lograr un resultado utilizando su voz o un chatbot. El sistema automatiza las tareas asociadas con el control de plazos de los recursos judiciales presentados; analiza la documentación pertinente que acompaña al expediente y con un sistema basado en consultas, de solo cinco preguntas, los jueces pueden desarrollar una opinión legal para decidir sobre un recurso.
- Automatización: el concepto de automatización tiene diferentes sutilezas basadas en numerosas circunstancias. Principalmente hay dos grandes grupos:

133 Lee J. (2020). Legal Tech-ing Our Way to Justice, disponible en: https://lawtech.asia/legal-tech-ing-our-way-to-justice/. Véase también: https://www.a-star.edu.sg/docs/librariesprovider10/default-document-library/fw-new-infosheets/smart-nation-digital-economy/intelligent-court-transcription-system.pdf?sfvrsn=72a5a971_3

134 Cátedra UNESCO de Sociedades del Conocimiento y Gobierno Digital (2020). PROMETEA: Transforming the administration of justice with artificial intelligence tools, disponible en: https://unescochair.cs.uns.edu.ar/en/2020/06/prometea-transforming-the-administration-of-justice-with-artificial-intelligence-tools/. Véase también: Corvalan J. G., Le Fevre Cervini E. M. (2020). Experiencia Prometea. Using AI to optimize public institutions, disponible en: https://ceridap.eu/prometea-experience-using-ai-to-optimize-public-institutions; https://www.ibanet.org/article/14AF564F-080C-4CA2-8DDB-7FA909E5C1F4

135 Corvalan J. G., Le Fevre Cervini E. M. (2020). Experiencia Prometea. Using AI to optimize public institutions, disponible en: https://ceridap.eu/prometea-experience-using-ai-to-optimize-public-institutions

- Automatización completa: los algoritmos asocian automáticamente datos e información con documentos. El documento se genera sin interacción de una persona.
- Automatización con intervención humana reducida: en muchos casos, se requiere la interacción humana con un sistema automatizado para completar o mejorar la generación de un documento.

- Clasificación y detección inteligente: la detección se basa en la lectura y el análisis de un volumen masivo de información, en el que Prometea puede identificar documentos en función de diferentes combinaciones de criterios, independientemente de la diversidad lingüística de los mismos. Luego, el sistema segmenta los datos en función de patrones compartidos (palabras clave) a lo largo de los documentos.
- Predicción: es la función más compleja que ofrece Prometea. Se hará una predicción basada en respuestas anteriores. Cuando Prometea encuentra una coincidencia entre el presente documento y uno anterior, toma nota de la respuesta proporcionada en situaciones anteriores y sugiere el mismo remedio porque las condiciones son similares. Este trabajo se deriva de la lectura y el reconocimiento de patrones de decisiones judiciales precedentes accesibles en la web de instancias anteriores. Una vez que Prometea identifica la solución, permite a la persona usuaria completar la opinión legal en función de algunas preguntas y luego muestra una vista previa editable en línea del documento final. El primer borrador del documento es generado automáticamente por el sistema de IA.[136]

Dadas las continuas preocupaciones con respecto a la justificación de las decisiones de Prometea y sus consecuencias para el debido proceso, la sociedad civil ha instado a una supervisión sostenida de la ejecución del programa. Otros aspectos a tener en cuenta son el nivel de responsabilidad de los actores relevantes (desarrolladores y jueces) y los posibles sesgos en los datos y el diseño del entrenamiento.[137]

PretorIA, Colombia

A principios de 2019, la Corte Constitucional colombiana anunció un proyecto piloto de implementación de Prometea para resolver la ineficiencia y los retrasos. Todos los días, la Corte recibe más de 2.000 órdenes de protección de todos los tribunales de todo el país. Solo nueve jueces y menos de 200 miembros del personal trabajan para la Corte Constitucional. Sin embargo, académicos y miembros de la sociedad

136 Ibid.
137 OECD, AI use cases in LAC governments, disponible en: https://www.oecd-ilibrary.org/sites/08955f48-en/index.html?itemId=/content/component/08955f48-en

civil plantearon numerosas preocupaciones sobre los posibles efectos de Prometea, así como su funcionamiento y el proceso de toma de decisiones que se consideraron opacos. Prometea resultó ser un piloto que se ha puesto en espera. El mayor desafío estaba relacionado con la privacidad y la protección de datos relacionados con el intercambio de información confidencial con terceros, como los desarrolladores de software. Es crucial que la identidad de las víctimas y su información o datos personales estén protegidos en los casos en que estén involucrados menores de edad, o en los casos en que estén implicados delitos sexuales, entre otras circunstancias. El acceso a esta información o datos por parte de cualquier persona que no sea el tribunal y las partes involucradas en el procesamiento de casos constituyó una violación de la confidencialidad. Dada la debilidad del sistema a este respecto, era especialmente preocupante que pudiera ocurrir una posible fuga de información personal a los medios de comunicación u otras partes interesadas, con resultados potencialmente desastrosos para la protección de la privacidad de las personas involucradas en los casos procesados por el sistema de IA.[138]

Tras múltiples debates, la Corte Constitucional cambió el proyecto implementando una tecnología más clara y transparente. PretorIA, lanzado a mediados de 2020, utiliza tecnología de modelado de temas en lugar de redes neuronales debido a esto. La nueva versión se puede explicar, interpretar y rastrear por completo.[139]

SUPACE, India

El Poder Judicial de la India tiene un gran número de casos pendientes. Según los datos de National Judicial Data Grid, alrededor de 38 millones de casos están pendientes en varios tribunales de distrito y de taluka en la India, y más de cien mil casos han estado pendientes durante más de tres décadas.[140]

La Corte Suprema de la India ha implementado un sistema de IA, el Portal de la Corte Suprema para la asistencia en la eficiencia de los tribunales (SUPACE) que ayudará en la administración e impartición de justicia a través de la catalogación de un gran número de decisiones judiciales anteriores para un mejor procesamiento del material del caso, ya sea para comprender la matriz de hechos de instancias específicas o para realizar una investigación dinámica de los precedentes. SUPACE no se utilizará en la toma de decisiones. El papel de la IA se limitará a la recopilación y el análisis de datos.[141]

138 Guitierrez O. L. C., Castañeda J. D., Saavedra Rionda V. P. (2019). Enthusiasm and complexity: Learning from the "Prometea" pilot in Colombia's judicial system, disponible en: https://giswatch.org/node/6166
139 Ibid.
140 Shanthi S. (2021). Behind SUPACE: The AI Portal Of The Supreme Court of India, disponible en: https://analyticsindiamag.com/behind-supace-the-ai-portal-of-the-supreme-court-of-india/
141 Ibid.

La herramienta SUPACE AI se está implementando de forma experimental con jueces que manejan casos penales en los Tribunales Superiores de Bombay y Delhi.

La Corte Suprema de la India está explorando el uso de una aplicación móvil que traducirá sus decisiones a nueve idiomas. Además, la India está utilizando la IA para resolver cargos menores, como infracciones de tráfico.[142]

HART (Herramienta de evaluación de riesgos), Reino Unido

La Herramienta de evaluación de riesgos (HART) es utilizada por la Policía de Durham en el Reino Unido. Utilizando más de treinta características que describen los antecedentes penales y socioeconómicos de una persona, HART utiliza un algoritmo de aprendizaje automático para determinar la probabilidad de reincidencia de un sospechoso. La policía local utiliza las evaluaciones de riesgo completadas por HART para decidir si se acusa a una persona o se remite a un programa de rehabilitación. HART no decide si una persona es culpable o inocente, pero su evaluación puede iniciar una serie de acciones que conducen a que una persona sea privada de su libertad o sea declarada culpable de un delito. Sin duda, los cargos deberían estar determinados por los méritos de cada caso individual, y es difícil ver cómo los juicios sobre la participación en programas de rehabilitación podrían decidirse de otra manera que analizando cuidadosamente la situación concreta de cada persona. Siempre debe haber un ser humano en el circuito que supervise el resultado de un sistema automatizado de toma de decisiones que tome decisiones de alto impacto y sensibles a los hechos.[143]

HART es propenso a criminalizar en exceso, ya que está intencionalmente destinado a subestimar quién está calificado para la vinculación en el programa de rehabilitación. Este método va en contra de la idea de que toda ambigüedad en un caso penal debe resolverse a favor del acusado (“in dubio reo”). Contrariamente a lo que hace HART, un enfoque de cumplimiento de los derechos humanos en la toma de decisiones de la justicia penal tendría que favorecer al acusado.[144]

142 Ibid.
143 Oswald M., Grace J., Urwin S., Barnes G. C. (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and ‘Experimental’ proportionality, Information & Communications Technology Law, 27 (2), 223–250, disponible en: https://doi.org/10.1080/13600834.2018.1458455
144 Ibid.

3. Actividades

Las siguientes actividades grupales están orientadas a alentar a las personas participantes de la capacitación a analizar diversas implicaciones relacionadas con el uso de la IA en el poder judicial

Actividad 1

Debata las siguientes preguntas con otras personas participantes de la capacitación:

- ¿Quién debe ser responsable de las decisiones automatizadas y cómo se debe asignar la responsabilidad dentro de la cadena de actores cuando la IA facilita la decisión final?
- ¿Qué es un juicio justo si la ADM ha facilitado las decisiones?
- ¿Se le niega al acusado el debido proceso legal cuando se implementan sistemas de IA en alguna etapa del procedimiento penal?

Actividad 2

Vaya al siguiente enlace: https://bja.ojp.gov/program/psrac/basics/what-is-risk-assessment#illustration. La ilustración "demuestra cómo se calculan las puntuaciones de riesgo en la evaluación de riesgos. En aras de la ilustración, este ejemplo hipotético solo abarca cinco dominios de predictores, que incluyen datos demográficos, antecedentes penales, educación/empleo, apoyo familiar/social y cognición antisocial, y solo un indicador para cada dominio. A los valores de cada indicador se les han asignado puntajes que van de 0 a 2; cuanto mayor sea el puntaje, más probabilidades tendrá de reincidir (por ejemplo, porque las personas más jóvenes tienen más probabilidades de reincidir que las personas mayores, los valores del indicador de "edad al momento de la sentencia" disminuyen a medida que aumenta la edad)".[145]

Introduzca ciertas características para comprender mejor cómo funciona la herramienta de evaluación de riesgos. Debata con otras personas participantes de la capacitación cuáles son sus ventajas y desventajas.

Actividad 3

Las personas participantes de la capacitación leen el escenario hipotético: "Navegando los riesgos: los jueces usan la IA generativa" y analizan los desafíos clave en el despliegue de la IA generativa por parte de los tribunales.

145 https://bja.ojp.gov/program/psrac/basics/what-is-risk-assessment

Descripción del escenario:

En un futuro en el que la IA generativa ha logrado avances significativos, los jueces han comenzado a experimentar con su uso en la sala de audiencias. Sin embargo, pronto se encuentran con varios desafíos y riesgos asociados con su adopción. Este escenario destaca los riesgos y peligros potenciales del uso de la IA generativa en un contexto judicial.

Elementos del escenario:

1. **Generación automatizada de documentos legales:**
   - Los jueces comienzan a usar IA generativa para automatizar la redacción de documentos legales, como sentencias y opiniones.
   - El sistema de IA, aunque eficiente, a veces genera argumentos y conclusiones legales sesgados o inexactos.

2. **Dependencia excesiva de la asistencia de IA:**
   - Los jueces confían cada vez más en el análisis legal generado por la IA, reduciendo gradualmente sus propias habilidades de pensamiento crítico y toma de decisiones.
   - Existe una creciente preocupación de que los jueces puedan convertirse en usuarios pasivos de la IA, disminuyendo su papel en la interpretación y aplicación de la ley.

3. **Sesgo ético y legal:**
   - Los modelos de IA utilizados por los jueces heredan sesgos presentes en sus datos de entrenamiento. Esto conduce a decisiones que favorecen desproporcionadamente a ciertos grupos o perpetúan los sesgos existentes en el sistema legal.
   - Los juristas y activistas plantean preocupaciones sobre la equidad y la discriminación.

4. **Transparencia y responsabilidad:**
   - Los modelos generativos de IA pueden ser complejos y difíciles de interpretar. Los jueces enfrentan desafíos para explicar las decisiones generadas por la IA a los litigantes, los abogados y el público.
   - Surgen preguntas sobre la responsabilidad de las decisiones generadas por la IA, particularmente en los casos en que tienen consecuencias negativas.

5. **Protección de datos y seguridad:**
   - El uso de IA generativa en los procedimientos judiciales implica el manejo de grandes cantidades de datos legales confidenciales. Surgen preocupaciones sobre las filtraciones de datos y la seguridad de la información confidencial.

- Los tribunales deben invertir mucho en ciberseguridad para protegerse frente a posibles amenazas.

6. **Confianza y percepción públicas:**
   - A medida que la IA generativa se vuelve más integral en el proceso legal, la confianza pública en el sistema de justicia se erosiona.
   - Los ciudadanos y los litigantes expresan escepticismo sobre la equidad e imparcialidad de las decisiones asistidas por IA.

7. **Desafíos y precedentes legales:**
   - Surgen desafíos legales sobre la admisibilidad de las pruebas generadas por IA y si la IA puede considerarse una fuente confiable de análisis legal.
   - Los tribunales se enfrentan a la tarea de establecer precedentes legales que rijan el uso de la IA en sus decisiones.

Resultado del escenario:

A medida que los jueces lidian con los riesgos y desafíos asociados con el uso de la IA generativa en la sala del tribunal, deben equilibrar cuidadosamente los beneficios potenciales de la eficiencia y la precisión con la necesidad de preservar la transparencia, la imparcialidad y el juicio humano en el sistema legal. El escenario subraya la importancia de directrices integrales, mecanismos de supervisión y capacitación continua para mitigar estos riesgos y garantizar que la IA mejore, en lugar de socavar, los principios de justicia.

4. Recursos

1. AAAS, Artificial Intelligence and the Courts: Materials for Judges, disponible en: https://www.aaas.org/ai2/projects/law/judicialpapers
2. Abu Elyounes D. (2019). Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness, Journal of Law, Technology and Policy, disponible en: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3478296
3. Ada Lovelace Institute, AI Now Institute y Open Government Partnership (2021). Algorithmic Accountability for the Public Sector, disponible en: https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/
4. Bhuiyan J. (2021). El LAPD puso fin a los programas de vigilancia predictiva en medio de la protesta pública. A new effort shares many of their flaws, disponible en: https://www.theguardian.com/us-news/2021/nov/07/lapd-predictive-policing-surveillance-reform
5. Brittain B. (2023). Getty Images lawsuit says Stability AI misused photos to train AI, disponible en: https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/Council of Europe, available at: https://www.coe.int/en/web/cepej
6. Consejo de Europa (2021). Plan de acción de la CEPEJ 2022 – 2025: "Digitalización para una justicia mejor", disponible en: https://rm.coe.int/cepej-2021-12-en-cepej-action-plan-2022-2025-digitalization-justice/1680a4cf2c
7. Futurism (2023). CNET's AI Journalist Appears to Have Committed Extensive Plagiarism, disponible en: https://futurism.com/cnet-ai-plagiarism
8. Hind M. (2019). Explaining Explainable AI por Michael Hind, The ACM Magazine for Students, 25(3), disponible en: https://doi.org/10.1145/3313096
9. Jauhar A., Misra M., Sengupta A., Chakrabarti P. P., Ghosh S., Ghosh K., (2021). Responsible Artificial Intelligence for the Indian Justice System, disponible en: https://vidhilegalpolicy.in/wp-content/uploads/2021/04/Responsible-AI-in-the-Indian-Justice-System-A-Strategy-Paper.pdf
10. IT world Canada (2023). Microsoft, GitHub, and OpenAI ask court to dismiss AI copyright lawsuit, disponible en: https://www.itworldcanada.com/post/microsoft-github-and-openai-ask-court-to-dismiss-ai-copyright-lawsuit
11. Somepalli G., Singla V., Goldblum M., Geiping J., Goldstein J. (2022). Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models, Universidad de Maryland, disponible en: https://arxiv.org/pdf/2212.03860.pdf
12. The Surveillance and Policing of Looted Land (2021). Automating banishment, disponible en:https://automatingbanishment.org/section/2-architecture-of-data-driven-policing/
13. MOOC de la UNESCO sobre IA y el Estado de derecho, disponible en: https://www.unesco.org/en/articles/unesco-global-mooc-ai-and-rule-law-engaged-thousands-judicial-operators
14. UNESCO (2021). Global Toolkit for Judicial Actors: International legal standards on freedom of expression, access to information and safety of journalists, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000378755

15. Vincent J. (2022). The lawsuit that could rewrite the rules of AI copyright, disponible en: https://www.theverge.com/2022/11/8/23446821/microsoft-openai-github-copilot-class-action-lawsuit-ai-copyright-violation-training-data

16. Vincent J. (2023). AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit, disponible en: https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart

17. Završnik A. (2019). Algorithmic justice: Algorithms and big data in criminal justice settings, European Journal of Criminology, 18, 623–642, disponible en: https://doi.org/10.1177/1477370819876762

Módulo 3

Desafíos legales y éticos de la IA

El módulo tres analiza los riesgos legales y éticos asociados con los sistemas de IA, y los desafíos de la transparencia algorítmica y la rendición de cuentas en el poder judicial. Luego procede con una descripción general de los problemas legales más destacados relacionados con la identificación biométrica y la tecnología de reconocimiento facial. El módulo también profundiza en los desafíos clave relacionados con la IA y la ética basados en la Recomendación de la UNESCO 2021 sobre la Ética de la Inteligencia Artificial.

¿Qué va a aprender?

Después de completar este módulo, las personas participantes podrán:

- Comprender y explicar los desafíos clave relacionados con la transparencia algorítmica y la rendición de cuentas en el poder judicial, y la jurisprudencia pertinente;
- Comprender los problemas legales más destacados relacionados con la identificación biométrica, la tecnología de reconocimiento facial y las falsedades;
- Tener una comprensión firme de los desafíos clave relacionados con la IA y la ética basados en la Recomendación de la UNESCO sobre la Ética de la Inteligencia Artificial (2021).

1. ¿Qué es la ética de la IA?

La Recomendación de la UNESCO sobre la ética de la IA aborda la ética de la IA como una reflexión normativa sistemática, basada en un marco holístico, integral, multicultural y en evolución de valores, principios y acciones interdependientes que pueden guiar a las sociedades a tratar de manera responsable los impactos conocidos y desconocidos de las tecnologías de IA en los seres humanos, las sociedades y el medio ambiente y los ecosistemas, y les ofrece una base para aceptar o rechazar las tecnologías de IA.

La UNESCO considera la ética como una base dinámica para la evaluación normativa y la orientación de las tecnologías de la IA, refiriéndose a la dignidad humana, el bienestar y la prevención de daños como brújula y enraizada en la ética de la ciencia y la tecnología.

En la práctica, la IA ética implica considerar las implicaciones éticas de los sistemas de IA y garantizar que su diseño e implementación se alineen con los valores y normas sociales más amplios.

Experimento mental:

Intentemos un experimento mental: están en la parada del tranvía y de repente notan un tranvía acelerando hacia cinco personas que no son conscientes de su aproximación. También ven una segunda pista que solo tiene una persona. ¿Qué harían? ¿Elegirían desviar el tranvía a la segunda vía para salvar a las cinco personas a costa de una vida?

Durante muchos años, el problema del tranvía ha sido un dilema ético de renombre abordado en los cursos de filosofía. Sin embargo, la aparición de autos autónomos experimentales ha convertido en realidad este problema teórico. Como resultado, ahora nos enfrentamos al desafío de determinar la programación adecuada para los sistemas de IA en situaciones críticas de vida o muerte.

Fuente: Universidad de Utrecht, Unboxing the black box of AI, disponible en: https://www.uu.nl/en/organisation/in-depth/unboxing-the-black-box-of-ai

Muchas iniciativas de autorregulación se han centrado en los riesgos éticos que plantea la IA. Los gobiernos, las organizaciones internacionales, el sector privado y las organizaciones de la sociedad civil han elaborado normas y principios éticos no vinculantes para guiar el desarrollo y el uso de la IA. Este capítulo ofrece una visión general de los marcos éticos clave de la IA, centrándose en la Recomendación de la UNESCO sobre la ética de la inteligencia artificial (2021). Es importante tener en cuenta que la Recomendación de la UNESCO, así como otros marcos éticos sobre la IA, no tienen los efectos vinculantes de la ley.

La Tabla 4 a continuación ofrece una descripción general de los principios fundamentales de la ética de la IA.

Tabla 4. Principios fundamentales de ética de la IA

Principios	Explicación
Equidad y sesgo	Los sistemas de IA deben diseñarse para garantizar la equidad y evitar sesgos que puedan conducir a resultados discriminatorios. Es crucial abordar los sesgos en los datos de entrenamiento, los algoritmos y los procesos de toma de decisiones para evitar el trato injusto o la marginación de ciertas personas o grupos.
Transparencia y explicabilidad	Los sistemas de IA deben ser transparentes y proporcionar a las personas usuarias una comprensión sobre cómo funcionan y cómo se toman las decisiones. La explicabilidad es importante para garantizar la rendición de cuentas, permitir la auditoría y generar confianza en las tecnologías de IA.
Política de privacidad y protección de datos	Los sistemas de IA a menudo se basan en grandes cantidades de datos, incluida información personal y confidencial. Respetar los derechos de privacidad y cumplir con las regulaciones de protección de datos son esenciales en el desarrollo y la implementación de la IA. Minimizar la recopilación de datos, garantizar el consentimiento informado y proteger los datos del acceso no autorizado son consideraciones fundamentales. Respetar, proteger y promover la privacidad es muy importante para salvaguardar la dignidad humana, la autonomía y la agencia a lo largo de todo el ciclo de vida de los sistemas de IA.[146]
Obligaciones y responsabilidad	Se deben establecer líneas claras de responsabilidad para los resultados de los sistemas de IA, incluida la identificación de quién es responsable de las acciones y decisiones tomadas por las tecnologías de IA. Es crucial garantizar que existan mecanismos para reparar los posibles impactos negativos de los sistemas de IA.
Seguridad y robustez	Los sistemas de IA deben diseñarse teniendo en cuenta la seguridad para evitar daños no deseados. Se deben tomar medidas para garantizar que las tecnologías de IA sean sólidas, confiables y capaces de manejar circunstancias imprevistas y ataques antagónicos.
Autonomía humana y supervisión	La IA debe desarrollarse y utilizarse para mejorar la autonomía humana y la toma de decisiones, en lugar de reemplazar o influir indebidamente en el juicio humano. Mantener la supervisión y el control humanos sobre los sistemas de IA es importante para preservar la actividad humana. Es crucial asegurarse de que la responsabilidad ética y legal se pueda asignar a personas físicas o entidades legales existentes en cada etapa del ciclo de vida del sistema de IA. Esto incluye los casos en los que se necesitan recursos. La supervisión humana significa algo más que la supervisión individual; también implica un monitoreo público inclusivo según sea necesario.[147]
Repercusiones sociales, ambientales y económicas	Las tecnologías de IA pueden tener profundos impactos sociales y económicos. Las consideraciones éticas incluyen garantizar un acceso equitativo a los beneficios de la IA, minimizar el desplazamiento laboral y abordar implicaciones sociales más amplias, como la desigualdad de la riqueza y la brecha digital.

146 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000380455
147 Ibid.

Principios	Explicación
Inclusividad y diversidad	Es crucial priorizar el respeto, la protección y la promoción de la diversidad y la inclusión al desarrollar sistemas de IA, de conformidad con el derecho internacional y los derechos humanos. Esto se puede lograr fomentando la participación activa de todas las personas y grupos, independientemente de su raza, color, ascendencia, género, edad, idioma, religión, opiniones políticas, origen nacional o étnico, origen social o económico, discapacidad o cualquier otro factor.[148]
Colaboración y enfoques multidisciplinarios	Abordar la ética de la IA requiere la colaboración entre varias partes interesadas, incluidos investigadores, responsables políticos, expertos de la industria, especialistas en ética y la sociedad civil. Las perspectivas multidisciplinarias y las voces diversas son cruciales para navegar por los complejos desafíos éticos de la IA.

¿Quién es un "actor de IA"?

De acuerdo con la Recomendación de la UNESCO sobre la Ética de la Inteligencia Artificial (2021), cualquier actor involucrado en al menos una etapa del ciclo de vida del sistema de IA se denomina "actor de IA". Esto incluye tanto a personas físicas como jurídicas, incluidos investigadores, programadores, ingenieros, científicos de datos, personas usuarias finales, empresas comerciales, instituciones académicas y entidades públicas y privadas.

¿Qué tipo de preocupaciones éticas plantean los sistemas de IA?

Los sistemas de IA plantean nuevas preocupaciones éticas, como las relacionadas con la toma de decisiones, el empleo y el trabajo, la interacción social, la atención médica, la educación, los medios de comunicación, el acceso a la información, la brecha digital, los datos personales y la protección del consumidor, la igualdad de género, el medio ambiente, la democracia, el Estado de derecho, la seguridad y la policía, el doble uso y los derechos humanos y las libertades fundamentales, como el derecho a la privacidad[149] la libertad de expresión y la igualdad ante la ley.

Además, el potencial de los algoritmos de IA para reproducir y reforzar prejuicios preexistentes e intensificar las formas existentes de discriminación, prejuicios y estereotipos presenta importantes desafíos éticos. A largo plazo, los sistemas de IA pueden socavar el valor agregado previamente asegurado a través del sentido único de agencia y experiencia de los humanos, lo que plantea nuevas preguntas sobre la autoconciencia humana, las interacciones sociales, culturales y ambientales, así como la autonomía, la agencia, el valor y la dignidad.[150]

148 Ibid.

149 Un documento notable en el área de cuestiones de privacidad y protección de datos para el poder judicial son las Directrices de la UNESCO (2022) para actores judiciales sobre privacidad y protección de datos, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000381298

150 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000380455

Actividad: ¿La IA toma mejores decisiones que los humanos? Pensar en la ética de la IA

Los participantes de la capacitación ven el video y debaten cómo interactúan la IA y la ética y cuál es el impacto de la IA en la ética y los derechos humanos.

Fuente: UNESCO, https://youtu.be/2E7l1hdjHsg

Marcos clave para la ética de la IA

Además de la Recomendación de la UNESCO sobre la Ética de la Inteligencia Artificial (2021), a continuación se presentan brevemente algunos otros marcos:

- La Iniciativa Global del IEEE sobre Ética de los sistemas autónomos e inteligentes: la Asociación de Normalización del IEEE ha elaborado una serie de documentos, como el marco de diseño éticamente alineado[151] y la serie de normas P7000[152]. Estos recursos proporcionan un enfoque integral de la ética de la IA, que abarca áreas como la transparencia, la rendición de cuentas y la priorización de los valores humanos.[153]
- Directrices éticas de la Comisión Europea para una IA confiable: la Comisión Europea publicó directrices que describen siete requisitos clave para una IA confiable: agencia y supervisión humanas, robustez y seguridad técnicas, privacidad y gobernanza de datos, transparencia, diversidad, no discriminación y bienestar social y ambiental.[154]

151 IEEE (2019). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (A/IS), disponible en:https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
152 Véase: https://sagroups.ieee.org/7000/
153 Véase: https://standards.ieee.org/industry-connections/ec/autonomous-systems/
154 Comisión Europea (2019). Ethics guidelines for trustworthy AI, disponible en: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

- El 3 de noviembre de 2017, se anunció la Declaración de Montreal para la IA Responsable al final del Foro sobre el Desarrollo Socialmente Responsable de la IA celebrado en Montreal. La Declaración es un ejemplo de un esfuerzo de cocreación que desarrolló un conjunto de principios rectores para el desarrollo y despliegue responsable de la IA para fines públicos. Este fue un esfuerzo de colaboración que involucró una serie de consultas públicas y asambleas ciudadanas con más de 500 residentes, expertos y partes interesadas clave. Con más de 2.200 ciudadanos y más de 200 organizaciones firmantes de la declaración, aboga por los siguientes principios: Bienestar, Privacidad e intimidad, Respeto a la autonomía, Responsabilidad, Participación democrática, Equidad, Solidaridad, Diversidad e inclusión, Prudencia y Desarrollo sostenible.[155]
- Los Principios de IA de Asilomar: estos principios fueron desarrollados por un grupo de investigadores, formuladores de políticas y pensadores de IA durante la Conferencia de Asilomar sobre IA beneficiosa. Abarcan diversos aspectos éticos, incluida la garantía de los amplios beneficios de la IA, la seguridad a largo plazo, el liderazgo en investigación técnica y la orientación cooperativa.[156]
- Los Principios de IA de la OCDE priorizan el desarrollo de IA confiable con un enfoque centrado en el ser humano. Elaborado con el aporte de un panel de más de 50 personas expertas que incluye gobiernos, academia, empresas, sociedad civil, organizaciones internacionales, la comunidad tecnológica y sindicatos, hay cinco principios centrados en valores para la implementación responsable y confiable de la IA, así como cinco recomendaciones para políticas públicas y colaboración global. Su objetivo es proporcionar orientación a los gobiernos, organizaciones e individuos en el desarrollo y funcionamiento de sistemas de IA que prioricen el bienestar de las personas, y garantizar que los responsables de su funcionamiento asuman su responsabilidad.[157]

La Tabla 5 al comienzo del módulo cuatro ofrece una descripción general de las iniciativas clave sobre regulación, política y ética de la IA.

¿Cómo poner en práctica la ética de la IA?

Cualquier iniciativa de IA en el poder judicial debe adherirse a las normas éticas de responsabilidad y apertura. El IEEE recomienda crear nuevos estándares que especifiquen grados de transparencia cuantificables y comprobables para que los sistemas puedan evaluarse de manera imparcial y se pueda establecer el grado de cumplimiento para mantener la transparencia.

155 Véase: https://gouai.cidob.org/resources/montreal-declaration-for-a-responsible-development-of-artificial-intelligence/
156 Un documento notable en el área de cuestiones de privacidad y protección de datos para el poder judicial son las Directrices de la UNESCO
157 UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000380455

Sin embargo, debido a los procesos intrincadamente vinculados y en capas de la programación algorítmica, mantener la transparencia del algoritmo es cada vez más difícil.[158] Los principios de ética de datos codificados o los códigos de conducta, las evaluaciones de impacto ético y las evaluaciones de impacto en la privacidad, la capacitación ética para operadores judiciales y las juntas de revisión ética son algunos ejemplos de métodos de revisión ética que pueden permitir una mayor transparencia y responsabilidad en el uso de los sistemas de IA y ADM en el sistema de justicia.

En general, las evaluaciones de impacto en la privacidad permiten a las organizaciones y a los desarrolladores evaluar de manera eficiente los riesgos planteados (garantizar el cumplimiento de los requisitos de privacidad, identificar medidas de mitigación y clasificar de manera efectiva los impactos del uso de datos y algoritmos). También sería ideal adoptar un enfoque inclusivo de las partes interesadas que enfatice "la inclusión proactiva de las personas usuarias". Además, se debe tener en cuenta constantemente el contexto de la utilización de datos, lo que requiere la intervención humana y, ocasionalmente, la experiencia específica del contexto.[159]

2. ¿Qué es el sesgo de IA?

El sesgo de IA es una diferencia sistemática en el tratamiento de ciertos objetos, personas o grupos (por ejemplo, estereotipos, prejuicios o favoritismo) en comparación con otros mediante algoritmos de IA. El sesgo de la IA puede afectar la recopilación e interpretación de datos, el diseño del sistema y la forma en que las personas usuarias interactúan con un sistema.[160]

Los sistemas de IA están lejos de ser elementos neutrales de tecnología. En cambio, pueden reflejar las preferencias, prioridades y prejuicios (inconscientes) de sus creadores. Los sesgos pueden surgir de muchas maneras en los sistemas de IA. Los datos de entrenamiento y los modelos de IA pueden estar sesgados. Los grupos privilegiados pueden tener ventajas en comparación con otros grupos en las decisiones de IA.

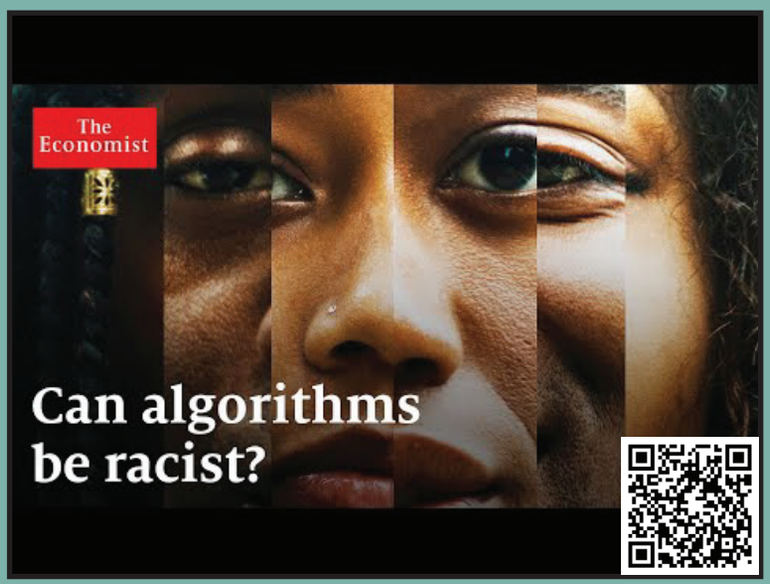
Incluso cuando los desarrolladores de software tienen mucho cuidado de minimizar cualquier influencia de su propio sesgo, los datos utilizados para entrenar un algoritmo pueden ser otra fuente importante de sesgo. Los sistemas de IA pueden reforzar lo que han aprendido de los datos y aumentar riesgos como el sesgo racial y de género.[161]

Además, incluso un algoritmo cuidadosamente construido debe basar sus juicios en información de una realidad impredecible e imperfecta. Los programas de IA son susceptibles de cometer errores de juicio en situaciones novedosas.[162]

158 Véase: https://www.ieee.org
159 Morley J., Floridi L., Kinsey L., Elhalal A. (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices, Eng Ethics, 26, 2141–2168, disponible en: https://doi.org/10.1007/s11948-019-00165-5
160 Goole (2023). Glosario de aprendizaje automático, disponible en: https://developers.google.com/machine-learning/glossary/
161 Turner Lee N., Resnick P., Barton G (2019). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms, disponible en: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/
162 Judge Dixon H. B. (2021). Artificial Intelligence: Benefits and Unknown Risks, disponible en: https://www.americanbar.org/groups/judicial/publications/judges_journal/2021/winter/artificial-intelligence-benefits-and-unknown-risks/

Punto de debate:

Los participantes de la capacitación ven el video y debaten cómo les ha afectado el sesgo de la IA y por qué es importante tenerlo en cuenta en los entornos judiciales



Fuente: The Economist, https://www.youtube.com/watch?v=IzvgEs1wPFQ

Experimento mental: sesgos basados en datos en la identificación de gatos y perros

Imagine que está creando un programa de IA para reconocer a las mascotas. Si el algoritmo está entrenado en un millón de imágenes de perros, pero solo en unos pocos miles de imágenes de gatos, puede tener dificultades para identificar con precisión a los gatos debido a una comprensión menos desarrollada de su apariencia. Vale la pena señalar que la IA puede exhibir sesgos, ya que se basa en datos y opciones de entrenamiento que pueden estar influenciados por sesgos humanos.

Fuente: Universidad de Utrecht, Unboxing the black box of AI, disponible en: https://www.uu.nl/en/organisation/in-depth/unboxing-the-black-box-of-ai

Algunos de los sesgos más controvertidos en la IA se producen en la tecnología de reconocimiento facial. Un estudio de 2016 realizado en Oakland, California, encontró que a pesar de que los datos de la encuesta muestran una distribución uniforme del consumo de drogas entre los grupos raciales, las predicciones algorítmicas de arresto policial se concentraron en comunidades predominantemente afroamericanas, creando bucles de retroalimentación que reforzaron los patrones de sesgo sistémico en la historia de los arrestos policiales.[163] Los algoritmos también pueden introducir sesgos raciales cuando los algoritmos de reconocimiento facial

163 Banco Mundial 2021. El estudio de 2016 realizado por el Grupo de análisis de datos de derechos humanos utilizando datos de 2010 y 2011 del departamento de policía de Oakland y otras fuentes comparó un mapeo del consumo de drogas basado en datos de encuestas de las víctimas de delitos con otro basado en el análisis algorítmico de las detenciones policiales. El estudio mostró que los datos de fuentes sesgadas podrían reforzar y potencialmente amplificar el sesgo racial en las prácticas de aplicación de la ley. Los datos sobre arrestos mostraron que los vecindarios afroamericanos tienen en promedio 200 veces más arrestos por drogas que otras áreas en Oakland.

se entrenan predominantemente con datos de rostros caucásicos, lo que reduce significativamente su precisión en el reconocimiento de otras etnias.[164] Es preocupante que varias tecnologías no funcionen con precisión para las personas con piel más oscura.

Por ejemplo, un estudio realizado por Georgia Tech ha revelado que los automóviles sin conductor tienen más probabilidades de golpear a las personas de color, ya que los sistemas de detección de objetos que utilizan para identificar a los peatones no funcionan tan eficazmente en personas con la piel más oscura. Estos ejemplos resaltan la necesidad de una tecnología más inclusiva e imparcial que cuide de todos, independientemente de su color de piel.[165] La industria tecnológica se ha enfrentado a un problema de larga data de diversidad en su fuerza laboral. El Informe de Salud de Internet 2020 de Mozilla sugiere que casi el 80 % de los empleados de los principales gigantes tecnológicos como Apple, Facebook, Google y Microsoft son hombres. Además, se ha registrado un crecimiento mínimo en la representación de las comunidades negras, latinas y nativas desde 2014, lo cual es una preocupación alarmante que debe abordarse.[166]

164 Hill K. (2020). Wrongfully Accused by an Algorithm." New York Times, disponible en: https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

165 Kenny C. (2021). Artificial Intelligence: Can We Trust Machines to Make Fair Decisions? Data and Computer Scientists, Ecologists, Pathologists, and Legal Scholars Study AI's Biases, disponible en: https://www.ucdavis.edu/curiosity/news/ais-race-and-gender-problem

166 Mozilla (2020). Internet Health Report, disponible en: https://foundation.mozilla.org/en/insights/internet-health-report/

En profundidad

Ejemplos de sesgo de IA

Microsoft Tay fue creado para atraer a personas de entre 18 y 24 años, y debutó en las redes sociales con un alegre "¡Hola, mundo!" (la "o" en "world" era un emoji del planeta Tierra). En doce horas, sin embargo, Tay se transformó en un negador del Holocausto racista y malhablado que declaró que todas las feministas "deberían morir y arder en el infierno". Tay, que se eliminó rápidamente de Twitter, fue diseñado para aprender de las acciones de otros usuarios de Twitter, y en este aspecto, tuvo éxito. La aceptación de Tay de las peores características de la humanidad es un ejemplo de sesgo algorítmico, que ocurre cuando un código aparentemente inofensivo adopta los sesgos de sus diseñadores o los datos que se alimentan.

En 2015, Google Photos identificó erróneamente a varios usuarios afroamericanos como gorilas, lo que provocó indignación en las redes sociales. El arquitecto social jefe de Google y jefe de infraestructura de Google Assistant, pronto anunció en Twitter que se estaba formando un equipo para abordar el problema.

Fuente: Wired (2017). How to Keep Your AI From Turning Into a Racist Monster, disponible en: https://www.wired.com/2017/02/keep-ai-turning-racist-monster/; véase también: https://www.bbc.com/news/technology-33347866.

Las herramientas de IA pueden estar sesgadas hacia las personas de color y las minorías

Una investigación de 2019 del Instituto Nacional de Ciencia y Tecnología de EE. UU. (NIST) sobre tecnologías de reconocimiento facial, que a menudo están "basadas en IA", descubrió que los algoritmos tenían hasta 100 veces más probabilidades de producir un falso positivo para las personas de color. Por ejemplo, el NIST descubrió que "para el emparejamiento de uno a muchos, el equipo vio mayores tasas de falsos positivos para las mujeres afroamericanas", un hallazgo que es "especialmente significativo porque las repercusiones pueden incluir acusaciones erróneas". La tasa de error para las mujeres de piel oscura fue del 34,7 %, pero la tasa de error para los hombres de piel clara fue del 0,8 %, según un segundo estudio realizado por la Universidad de Stanford y el MIT. Una evaluación de Rekognition, un sistema de reconocimiento facial propiedad de Amazon y vendido a las fuerzas del orden, descubrió indicadores de sesgo racial y descubrió que el sistema reconocía incorrectamente a 28 miembros del Congreso de los EE. UU. como delincuentes condenados. Del mismo modo, la IA y los sistemas algorítmicos de toma de decisiones empleados en las disposiciones previas al juicio, las sentencias y los contextos penitenciarios con frecuencia proporcionan resultados erróneos o sesgados que perpetúan las disparidades existentes.

Uno de los aspectos más desafiantes del sesgo de la IA es que los ingenieros y desarrolladores de IA no deben ser intencionalmente racistas o sexistas. Esta es una condición preocupante en un momento en que las personas creen cada vez más que la tecnología es más imparcial que ellos. A medida que la industria informática desarrolla la IA, corre el riesgo de incorporar el racismo y otros prejuicios en un código que tomará

decisiones durante décadas. Y debido a que el aprendizaje profundo implica que el código, no los humanos, escribirá código, la necesidad de eliminar el sesgo algorítmico es aún mayor.

El sesgo de IA puede ser generado por varias razones, y las siguientes son algunas definiciones y ejemplos de los principales sesgos de IA:

- **Sesgo de muestra debido a datos de entrenamiento sesgados y no representativos:** si las reglas extraídas por el algoritmo de aprendizaje automático de cualquier conjunto de datos se consideran legítimas, los prejuicios y omisiones incrustados en los datos de ejemplo se repetirán en el modelo predictivo. En otras palabras, si los datos utilizados para entrenar el modelo de IA no son representativos del contexto en el que se utilizará el sistema de IA, el sistema de IA puede producir resultados sesgados. Por ejemplo, un sistema de reconocimiento facial que se ha desarrollado predominantemente utilizando fotografías de hombres blancos, puede no ser capaz de identificar con precisión a las mujeres u otros grupos raciales. La investigación muestra que en el caso de las mujeres y las personas de diferentes orígenes raciales y culturales, los niveles de precisión de estos modelos son significativamente más bajos. Otro ejemplo serían los sistemas de IA programados para identificar el cáncer de piel. Si el conjunto de datos inicial no es representativo de la población, este método tendrá un rendimiento deficiente para los miembros de grupos subrepresentados.[167]
- **Sesgo de recuerdo durante el etiquetado de datos:** cuando la solución de IA utiliza datos etiquetados, el proceso de etiquetado debe ser coherente en todos los conjuntos de datos; de lo contrario, el resultado del modelo se vuelve inexacto. Por ejemplo, alguien podría describir una imagen de un teléfono como dañada, pero otra imagen comparable como ligeramente dañada. El conjunto de datos será inconsistente en esta situación, ya que habrá dos etiquetas diferentes que se refieren a imágenes similares y comparables.
- **Sesgo de asociación:** es importante tener en cuenta que incluso los conjuntos de datos representativos reflejan sesgos históricos y sociales, por ejemplo, contra las minorías excesivamente representadas en las poblaciones penitenciarias o las mujeres en trabajos menos prestigiosos. Por lo tanto, la "representatividad" de los datos puede perpetuar la discriminación y la desigualdad, cuando en realidad un conjunto de datos adaptado conscientemente que corrija tales desigualdades sociales podría producir resultados menos discriminatorios a partir de algoritmos entrenados sobre esta base y luego aplicados a nuevos casos (como

167 Cotton R. (2022). Data Demystified: The Different Types of AI Bias, disponible en: https://www.datacamp.com/blog/data-demystified-the-different-types-of-ai-bias

cuando se utiliza para informar la sentencia de custodia o el escrutinio automatizado de las solicitudes de empleo).El sesgo de asociación más conocido es el sesgo de género, como cuando el conjunto de datos utilizado se refiere a un grupo de profesiones donde todos los hombres trabajan como médicos y todas las mujeres como enfermeras. Esto no impide que los hombres se conviertan en enfermeras o que las mujeres se conviertan en médicos. Sin embargo, según el modelo de ML, no hay enfermeros ni doctoras.

- **Sesgo de medición:** es causado por una medición defectuosa por parte de los sujetos y/o el investigador. La fuente del sesgo de medición es una inexactitud que se produce durante la recopilación o medición de datos. Por ejemplo, si las fotos capturadas por una cámara utilizada para proporcionar datos para un sistema de reconocimiento de imágenes son de baja calidad, esto podría dar lugar a resultados sesgados frente a ciertos datos demográficos.[168] Otra ilustración es el juicio humano. Por ejemplo, un sistema de diagnóstico médico puede entrenarse para predecir la probabilidad de enfermedad en función de medidas indirectas, como visitas al médico, en lugar de síntomas reales.[169] El sesgo de medición también puede provenir de cuando los datos de ciertos grupos de población no se capturan en absoluto debido a su existencia fuera del flujo de recopilación de datos. Por ejemplo, el uso de datos de teléfonos móviles como indicador indirecto de la capacidad de la persona usuaria para pagar préstamos puede perjudicar a las personas con acceso limitado o nulo a los teléfonos móviles. Otro ejemplo sería una situación en la que un algoritmo diseñado para encontrar candidatos para trabajos potencialmente exitosos puede usar el éxito pasado en el lugar de trabajo como un predictor del éxito futuro en el lugar de trabajo y extraer de esa información criterios específicos de reclutamiento favorecidos como educación y experiencia. Sin embargo, las estadísticas subyacentes pueden estar desactualizadas, por ejemplo, en un momento en que las minorías o las mujeres estaban subrepresentadas en el mercado laboral relevante o en los estándares de admisión escolar. Como resultado, el sistema podría descalificar a los solicitantes que podrían superar el conjunto de datos de "desempeño laboral exitoso" del pasado.[170]
- **Sesgo de automatización debido a la dependencia acrítica de los resultados generados por la IA:** una amenaza importante que plantea el uso de sistemas de IA en la administración de justicia es el llamado sesgo de automatización, que es la tendencia de los humanos a considerar sin someter a crítica la solución ofrecida por la IA como correcta. Esto puede llevar a una falta de escepticismo hacia la información proporcionada por los algoritmos y una

168 Hackernoon (2020). 7 Types of Data Bias in Machine Learning, disponible en: https://hackernoon.com/7-types-of-data-bias-in-machine-learning-ubl3t3w.
169 Data Camp (2022). Different types of AI bias, disponible en: https://www.datacamp.com/blog/data-demystified-the-different-types-of-ai-bias.
170 Baker J. E., Hobart L. N., Mittelstead M. G. (2021). AI for Judges. A Framework. Centro de Seguridad y Tecnología Emergente, disponible en: https://www.armfor.uscourts.gov/ConfHandout/2022ConfHandout/Baker2021DecCenterForSecurityAndEmergingTechnology1.pdf

tendencia a actuar automáticamente en relación con lo que sugiere el algoritmo. Detectar el sesgo de automatización puede ser difícil, ya que a menudo es inconsciente. Una forma de detectarlo es prestar atención a cómo confiamos en la información proporcionada por los sistemas automatizados y sopesar si estamos siendo críticos con esa información o si la estamos aceptando sin cuestionarla. También es importante ser conscientes de nuestros propios sesgos y prejuicios y tratar de ser objetivos a la hora de evaluar la información proporcionada por los sistemas automatizados. Por lo tanto, la desviación del juez de cualquier decisión que sea asistida o automatizada no debe implicar ningún tipo de represalia, sanción, inspección o régimen disciplinario. Si prevalecen la supervisión y el control humanos, el control debe ser efectivo (consulte la sección sobre "El principio del humano en el circuito" en el Módulo 1).

Actividad: Los participantes en la capacitación leen la historia a continuación y evalúan el impacto ético de la tecnología siguiendo el instrumento de Evaluación de Impacto Ético de la UNESCO en el Anexo I [céntrese en las partes que se ocupan de la equidad, la no discriminación, la diversidad y la protección de datos y la privacidad]

En 2020, JK solicitó una licencia de conducir internacional en la Oficina Estatal de Transporte de Hamburgo, una ciudad portuaria del norte de Alemania. Presentó todo el papeleo requerido en su cita, excepto una foto biométrica ya que quería tomarla en la cabina de fotos de la oficina. Para tomar una foto biométrica, tenía que colocar su cara en un área específica de la cámara, y la foto solo se tomaría una vez que se detectara la cara allí. JK no fue reconocido en la cabina de fotos de la Oficina Estatal de Transporte, ya que parecía que solo las caras con tonos de piel claros eran reconocidas por el software de reconocimiento facial en esta cabina de fotos.

JK recordó que un miembro del personal le dijo que podría haber un problema con el tono de su piel. La oficina de impresión del gobierno era la propietaria de la cabina de fotos. Dijeron que, dado que la cabina de fotos estaba equipada con la tecnología más reciente, el problema no estaba relacionado con el software. En cambio, la oficina afirmó que la iluminación en la cabina era inadecuada y fue la causa del problema. Sin embargo, las tecnologías de IA más recientes pueden tener debilidades que resultan en resultados discriminatorios y sexistas, según un estudio realizado por Joy Buolamwini y Timnit Gebru.

Fuente: Algorithm Watch. Automated Decision-Making Systems and Discrimination Understanding causes, recognizing cases, supporting those affected A guidebook for anti-discrimination counselling; Buolamwini J., Gebru T. (2018). Sombras de género: disparidades interseccionales de precisión en la clasificación comercial de género. Actas de la 1.ª Conferencia sobre Equidad, Rendición de Cuentas y Transparencia, PMLR, 81, 77–91, disponibles en: https://proceedings.mlr.press/v81/buolamwini18a.html

¡Recordatorio!

Las herramientas de IA incorporan las elecciones de políticas de los tomadores de decisiones anteriores y, por lo tanto, el sesgo de esas decisiones. Las herramientas de juicio redactadas previamente, por ejemplo, pueden introducir prejuicios, reducir la discrecionalidad judicial y no abordar las dificultades específicas que enfrentan las personas de grupos marginados y vulnerables. Como resultado, comprender estas tecnologías y continuar investigándolas y evaluándolas asegurará que los jueces puedan participar plenamente en la evolución de las operaciones judiciales habilitadas por la IA.

Un llamado de atención: sobre el sesgo en los sistemas de IA que conducen a la discriminación

Incluso si un sistema de IA parece ser neutral en la superficie, sus algoritmos pueden llevar a evaluaciones y consecuencias discriminatorias. La discriminación a menudo puede surgir de prácticas prejuiciosas en el mundo real que alimentan los datos utilizados por el sistema de IA.

Cuando las tecnologías policiales basadas en datos son cajas negras, es difícil analizar los peligros de las tasas de error, los falsos positivos, las limitaciones en las capacidades de programación, los datos sesgados e incluso las fallas en el código fuente que tienen influencia en los resultados de búsqueda. Estos sistemas de cajas negras perpetúan ciclos viciosos de sesgo.

Los sistemas policiales predictivos que se basan en datos históricos corren el riesgo de replicar los resultados de actos discriminatorios anteriores. Esto puede dar lugar a "ciclos de retroalimentación", en los que cada nueva elección basada en datos anteriores genera más datos, lo que resulta en que los grupos marginados sean desproporcionadamente tomados como sospechosos y encarcelados. Los algoritmos predictivos pueden contribuir a la toma de decisiones sesgada y a las consecuencias discriminatorias dependiendo de cómo se documenten los delitos, qué delitos se eligen para incluirse en el estudio y qué métodos analíticos se emplean.

Aunque muchas personas creen que los datos policiales son neutrales, contienen sesgos políticos, sociales y de otro tipo. Los datos del departamento de policía reflejan los procedimientos y prioridades del departamento, así como los intereses locales, estatales y federales, y los prejuicios institucionales e individuales. No existen procedimientos definidos para utilizar la información recopilada durante las operaciones de aplicación de la ley en el desarrollo de sistemas de IA. Además, las prácticas policiales pueden tener una apertura y supervisión limitadas.[171]

171 Leslie D., Burr C., Aitken M., Cowls J., Katell M., Briggs M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer, The Council of Europe, disponible en: https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf. "Cada vez más, los gobiernos están adoptando regulaciones con respecto al uso de datos, como el Reglamento General de Protección de Datos (RGPD) de la Unión Europea. Pero estos tienden a centrarse en el uso corporativo de los datos y la medida en que las protecciones otorgadas en virtud de estas regulaciones se extienden a los AEL [agentes encargados de hacer cumplir la ley] es menos clara", disponible en: UNESCO (2022). Kit de herramientas global para agentes encargados de hacer cumplir la ley: libertad de expresión, acceso a la información y seguridad de los periodistas, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000383978

Muchos estudios de investigación han demostrado repetidamente que el uso de algoritmos predictivos en la vigilancia policial entrenados sobre datos delictivos pasados replica y amplifica los sesgos sistémicos existentes. A menudo, este proceso tiene poca consideración sobre cómo las diferentes iniciativas de reducción del delito, la legislación penal, las tendencias de elaboración de perfiles o los sesgos de sentencia influyen en los patrones detectados por dichos algoritmos en los datos.

El aumento del escrutinio público de estos algoritmos ha planteado preguntas sobre cómo se desarrollan e implementan; por qué no están sujetos a un mayor escrutinio; y si existen mecanismos de gobernanza para evaluar adecuadamente sus riesgos, vulnerabilidades y potencial para un daño social superior.[172] Se ha demostrado que el despliegue de herramientas de IA en el sistema de justicia penal puede exacerbar las prácticas policiales ya discriminatorias contra las minorías.

Actividad: La verdad sobre los algoritmos. Los participantes de la capacitación ven el video presentado por Cathy O'Neil y debaten cómo y por qué los algoritmos están sesgados. Los participantes también debaten cómo el sesgo algorítmico podría afectar su trabajo.

Fuente: https://youtu.be/heQzqX35c9A

Después de ver el video de capacitación, los participantes también analizan el siguiente escenario:

Escenario: sesgo algorítmico en la contratación

En un futuro no muy lejano, una gran corporación, llamémosla "TechCo", decide implementar un sistema de contratación algorítmico para agilizar su proceso de contratación y hacerlo más eficiente. TechCo se enorgullece de su compromiso con la diversidad y la inclusión, y la dirección cree que el uso de herramientas de contratación

172 Grupo de trabajo de la NACDL sobre vigilancia predictiva (2021). Garbage in, gospel out. How Data-Driven Policing Technologies Entrench Historic Racism and 'Tech-wash' Bias in the Criminal Legal System, disponible en: https://www.nacdl.org/getattachment/eb6a04b2-4887-4a46-a708-dbdaade82125/garbage-in-gospel-out-how-data-driven-policing-technologies-entrench-historic-racism-and-tech-wash-bias-in-the-criminal-legal-system-11162021.pdf

impulsadas por la IA le ayudará a alcanzar estos objetivos. Contratan a un equipo de científicos de datos e ingenieros de aprendizaje automático para desarrollar el sistema.

Así es como se desarrolla el escenario:

1. Recopilación de datos:
   - El equipo comienza recopilando datos históricos de los procesos de contratación anteriores de TechCo. Este conjunto de datos incluye currículums, comentarios de entrevistas y decisiones de contratación de la última década.
   - Infortunadamente, los datos históricos reflejan algunos sesgos que han existido dentro de la empresa. Por ejemplo, hay un número desproporcionado de candidatos masculinos contratados para roles técnicos, y los candidatos de ciertas universidades de prestigio son favorecidos.
2. Entrenamiento de modelos:
   - Los científicos de datos utilizan estos datos históricos para entrenar el algoritmo. Su objetivo es identificar patrones y criterios que predigan a los candidatos exitosos.
   - Debido a los datos históricos sesgados, el algoritmo comienza a detectar estos sesgos. Por ejemplo, podría aprender que los candidatos de universidades prestigiosas tienen más probabilidades de tener éxito, a pesar de que esta preferencia se basa en el sesgo histórico en lugar del mérito objetivo.
3. Sesgo involuntario:

   A medida que el algoritmo comienza a procesar nuevas solicitudes de empleo, perpetúa inadvertidamente los sesgos presentes en los datos de capacitación. Los currículums de mujeres, candidatos de entornos subrepresentados y aquellos de escuelas menos prestigiosas reciben puntajes más bajos, lo que lleva a su rechazo o a ser empujados al fondo del grupo de contratación.
4. Quejas y preocupaciones éticas:
   - Con el tiempo, los solicitantes de empleo que sienten que fueron rechazados injustamente comienzan a expresar sus preocupaciones. Notan un patrón en el que el algoritmo pone sistemáticamente en desventaja a ciertos grupos.
   - Las organizaciones de derechos civiles y los medios de comunicación se enteran de estos problemas y comienzan a investigar las prácticas de contratación de TechCo, acusándolas de sesgo algorítmico y discriminación.
5. Consecuencias legales y reputacionales:
   - TechCo se enfrenta a desafíos legales y posibles demandas por prácticas de contratación discriminatorias. También sufren un impacto significativo en su reputación, ya que los clientes y socios expresan su preocupación por su compromiso con la diversidad y la inclusión.
   - La dirección de la empresa se da cuenta del problema del sesgo algorítmico y decide detener temporalmente el uso del algoritmo de contratación mientras investiga el problema.
6. Auditoría algorítmica y medidas correctivas:
   - TechCo contrata auditores externos y especialistas en ética de datos para evaluar el algoritmo y su impacto. Los auditores identifican los datos sesgados y las fallas en el modelo.
   - La empresa toma medidas para volver a entrenar el algoritmo con un conjunto de datos más diverso y representativo, eliminar las características sesgadas e implementar salvaguardias contra sesgos futuros.
7. Reconstruir la confianza:

   TechCo se disculpa públicamente por el sesgo algorítmico y la discriminación. Describen su compromiso de rectificar el problema y garantizar prácticas de contratación justas.

La compañía invierte en medidas de transparencia, publicando regularmente informes sobre el desempeño de su algoritmo de contratación y buscando supervisión externa para recuperar la confianza.

Varios candidatos para el trabajo que creen que han sido perjudicados por el sesgo incrustado en el sistema de contratación presentan quejas. ¿Qué decidirán y qué factores tendrán en cuenta a la hora de tomar su decisión?

Los riesgos relacionados con los sesgos que plantean la IA y la ADM se han generalizado, como en los sistemas de reconocimiento facial en espacios públicos que permiten la vigilancia masiva[173] o en el despliegue de sistemas ADM muy sesgados para la detección de fraudes de bienestar, como el sistema holandés SyRI, que se analiza en el recuadro a continuación.[174] Los sistemas de IA pueden funcionar de manera impredecible, e incluso los sistemas que parecen realizar tareas "simples" o rutinarias pueden tener resultados no deseados y, a menudo, perjudiciales. Esto hace que los riesgos sean aún mayores, como se muestra en el recuadro a continuación que destaca ejemplos de sesgo algorítmico en el poder judicial.

Casos prácticos: ejemplos de sesgo algorítmico en el poder judicial y el gobierno

COMPAS

El Perfil de gestión de delincuentes correccionales para sanciones alternativas (COMPAS) utilizado por el poder judicial en los Estados Unidos no incluye la raza o el origen étnico como criterio, sin embargo, la investigación ha demostrado que asigna rutinariamente mayores puntajes de riesgo a los acusados negros que a los blancos, lo que hace que sea menos probable que sean liberados.[175] Ha habido casos en los que a presos con registros prácticamente perfectos, como Glen Rodríguez[176], se les ha negado la libertad condicional debido a una puntuación COMPAS inexacta, dejándolos con pocos recursos para impugnar la decisión o incluso averiguar cómo se calculó. Un análisis de 2016 realizado por ProPublica reveló que los COMPAS utilizados por los tribunales de Florida contenían prejuicios raciales. ProPublica examinó 7000 casos y descubrió que el puntaje era extraordinariamente poco confiable para predecir delitos violentos: solo el 20 % de los que se esperaba que cometieran delitos violentos lo hicieron. Los investigadores también descubrieron que era más probable que el algoritmo designara a los acusados de color como futuros delincuentes que a los acusados blancos, y que los acusados blancos eran etiquetados erróneamente con mayor frecuencia como de bajo riesgo que los acusados de color.[177] El propietario de COMPAS, Northpointe, publicó una réplica, que respondió al estudio de ProPublica y argumentó que el informe de ProPublica estaba "basado en estadísticas y análisis de datos defectuosos y no demostró que el COMPAS en sí mismo esté sesgado racialmente, y mucho menos que otros instrumentos de riesgo estén sesgados"[178].

173 Big Brother Watch (2019). UK MASS SURVEILLANCE CHALLENGED IN EUROPE'S HIGHEST HUMAN RIGHTS COURT, disponible en: https://bigbrotherwatch.org.uk/2019/07/uk-mass-surveillance-challenged-in-europes-highest-human-rights-court/

174 Algorithm Watch (2020). How Dutch activists got an invasive fraud detection algorithm banned, disponible en: https://algorithmwatch.org/en/syri-netherlands-algorithm/

175 Corbett-Davies S., Pierson E., Feller A., Goel S., Huq A. (2017). Toma de decisiones algorítmica y el costo de la equidad, disponible en: https://arxiv.org/pdf/1701.08230.pdf

176 Wexler R. (2017). When a computer program keeps you in jail: How computers are harming criminal justice, disponible en: https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

177 Criswell B. (2020). Algorithms Deciding the Future of Legal Decisions, disponible en: https://montrealethics.ai/algorithms-deciding-the-future-of-legal-decisions/

178 Angwin J., Larson J., Mattu S., Kirchner L. (2016). Machine Bias, disponible en: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; also see: https://www.uscourts.gov/federal-probation-journal/2016/09/false-positives-false-negatives-and-false-analyses-rejoinder.

El sistema SyRI

Para detectar el fraude en el bienestar social, el gobierno holandés implementó un sistema llamado SyRI, que significa por sus siglas en inglés "indicación de riesgo del sistema", para cruzar la información personal de los residentes de diferentes bases de datos e identificar "perfiles de ciudadanos poco deseables" que requieren un mayor escrutinio. El sistema funcionaba de la siguiente manera: si una agencia gubernamental (por ejemplo, municipios, el banco de seguridad social, las autoridades fiscales) detectaba fraude con beneficios, subsidios o impuestos en un determinado vecindario, podría usar SyRI. SyRI ayudó a identificar qué residentes debían ser investigados más a fondo por fraude.

A esta práctica se opusieron la Autoridad Holandesa de Protección de Datos y el Consejo de Estado, que plantearon preocupaciones sobre el derecho a la privacidad, así como los derechos al debido proceso, como la presunción de inocencia. Además, el sistema carecía de transparencia, ya que sus algoritmos no se publicaron y no se sometió a una auditoría técnica, y su focalización en los barrios desfavorecidos podría constituir una discriminación basada en la situación socioeconómica o migratoria de los residentes. Además, SyRI se ha utilizado sobre todo en barrios de bajos ingresos. Esto exacerba la discriminación y el sesgo si el gobierno utiliza exclusivamente el análisis de riesgos de SyRI en dichos barrios.

En 2020, el tribunal de La Haya ordenó[179] el cese inmediato de SyRI, por lo que concluyó que la legislación que establecía SyRI proporcionaba una protección insuficiente contra la intrusión en la vida privada, debido a las medidas desproporcionadas adoptadas para prevenir y castigar el fraude en interés del bienestar económico. El tribunal concluyó que SyRI violaba el artículo 8 del Convenio Europeo de Derechos Humanos (CEDH), que protege el derecho al respeto de la vida privada y familiar.

Fuente: Algorithm Watch (2020) How Dutch activists got an invasive fraud detection algorithm banned, disponible en: https://algorithmwatch.org/en/syri-netherlands-algorithm/. Véase también: https://towardsdatascience.com/fighting-back-on-algorithmic-opacity-30a0c13f0224; https://iapp.org/news/a/digital-welfare-fraud-detection-and-the-dutch-syri-judgment/; https://pace.coe.int/en/files/28715/html

Regulación del derecho a la explicación en la UE en el contexto de ADM

Reglas como el "derecho a la explicación" del Reglamento General de Protección de Datos (RGPD) de la UE se promulgaron en respuesta a problemas relacionados con la transparencia y la responsabilidad de la IA.[180] Los artículos 13(2)(f), 14(2)(g) y 15(1)(h) del RGPD obligan a los controladores de datos a informar a los interesados sobre la existencia de ADM, incluida la elaboración de perfiles, a la que se hace referencia en el artículo 22(1) y (4) e información significativa sobre la lógica involucrada, así como la importancia y las consecuencias para el interesado.[181]

179 Algorithm Watch (2020). How Dutch activists got an invasive fraud detection algorithm banned, disponible en: https://algorithmwatch.org/en/syri-netherlands-algorithm/.

180 Casey B., Farhangi A., Vogl R. (2018). Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise, Berkeley Technology Law Journal, 34, disponible en: https://ssrn.com/abstract=3143325

181 Un sujeto de datos es una persona que puede identificarse, ya sea directa o indirectamente, a través de un identificador como un nombre, número de identificación o datos de ubicación, o a través de factores personales relacionados con su identidad física, fisiológica, genética, mental, económica, cultural o social. Véase también: https://academic.oup.com/idpl/article/7/4/233/4762325

El artículo 22, apartado 1, del RGPD especifica que los interesados tienen derecho a no ser objeto de una decisión basada exclusivamente en el tratamiento automatizado, incluida la elaboración de perfiles, que genere efectos jurídicos que les conciernan o les afecte significativamente de manera similar. El artículo 22(2) – (4) describe las condiciones limitadas en las que se permite la toma de decisiones automatizada y describe ciertas protecciones para garantizar que los interesados puedan ejercer con éxito sus derechos.[182]

Caso de estudio: Legislación sobre el derecho a la explicación en Estonia

La Sección 23(4) de la Ley de Seguro de Desempleo de Estonia permite al Fondo de Seguro de Desempleo tomar decisiones sobre la asignación de beneficios de desempleo a los solicitantes de forma totalmente automática. Se informa inmediatamente a los solicitantes que la decisión se tomó automáticamente, que tienen derecho a ser escuchados y que pueden presentar una solicitud de revisión interna.

Dichas prácticas permiten a las personas cuya fecha ha sido sometida a la toma de decisiones automatizada comprender cómo se tomaron las decisiones y apelarlas.

Fuente: https://fpf.org/blog/gdpr-and-the-ai-act-interplay-lessons-from-fpfs-adm-case-law-report%ef%bf%bc/

Sesgo de IA e igualdad de género

Por ejemplo, las tecnologías de Reconocimiento automatizado de género (AGR) eliminan el derecho a la autoidentificación e infieren el género en función de los datos adquiridos sobre las personas. Las tecnologías AGR utilizan información como el nombre legal y las características faciales de una persona para simplificar la identidad de género a un binario. Esto carece de una comprensión científica de las diversas identidades de género.[183] Este borrado sistemático y tecnológicamente reforzado tiene efectos en el mundo real sobre los derechos fundamentales de las personas con diversas identidades de género y afecta el disfrute de sus derechos relacionados con la asistencia social, como la vivienda, el trabajo y los beneficios de salud.[184] Además, el diseño de conjuntos de datos puede afectar a la identidad de las personas. Un conjunto de datos que captura el género como binario, por ejemplo, confunde a las personas con diversas identidades de género.[185]

182 Ibid.

183 Sun S. D. (2019). Stop Using Phony Science to Justify Transphobia, disponible en: https://blogs.scientificamerican.com/voices/stop-using-phony-science-to-justify-transphobia/; see also UN, OHCHR and the human rights of LGBTI people, disponible en: https://www.ohchr.org/en/sexual-orientation-and-gender-identity. Véase también, UNESCO (2022). Glosario: Comprensión de conceptos en torno a la igualdad de género y la inclusión en la educación, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000380971

184 Leufer D. (2021). Computers are binary, people are not: how AI systems undermine LGBTQ identity, disponible en: https://www.accessnow.org/how-ai-systems-undermine-lgbtq-identity/

185 Consejo de Derechos Humanos de las Naciones Unidas (2021). El derecho a la privacidad en la era digital. The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, disponible en: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

AymurAI: Inteligencia artificial responsable para una justicia abierta y sensible al género

AymurAI es una iniciativa para promover la justicia abierta y sensible al género en América Latina. Esta iniciativa tiene como objetivo ayudar a los funcionarios de los tribunales penales y a los jueces que desean promover los datos abiertos en sus tribunales penales. AymurAI es un software basado en IA que identifica de forma semiautomática información importante en las resoluciones judiciales y crea conjuntos de datos abiertos centrados en los datos de violencia de género. También cuenta con una herramienta de anonimización que detecta información sensible en sentencias judiciales penales y la redacta. "AymurAI" significa "cosecha" en quechua. Esta herramienta tiene como objetivo "cosechar" datos de resoluciones judiciales en general, con un énfasis específico en los casos de violencia de género. Está "semiautomatizada" porque no funciona de forma autónoma sin la intervención humana y la toma de decisiones. AymurAI ayuda a detectar información relevante y agiliza la recopilación de sentencias judiciales, pero la validación humana de los hallazgos del software es crucial para garantizar resultados precisos y confiables.

AymurAI es una aplicación de escritorio que lee la resolución judicial, detecta información relevante, la presenta al usuario para su validación y luego la almacena en un conjunto de datos que se puede publicar. La herramienta utiliza reglas y Reconocimiento de entidad nombrada (NER) para extraer información esencial de documentos judiciales. En los casos de violencia de género, las etiquetas pueden representar el tipo de violencia, la ubicación, el género, la relación con el perpetrador, la decisión del juez en ese caso y otros datos relevantes. Estas etiquetas pasan por un proceso de validación y, una vez aprobadas, la información recopilada se estructura en un conjunto de datos abierto. Todo esto se consigue en cuatro sencillos pasos.

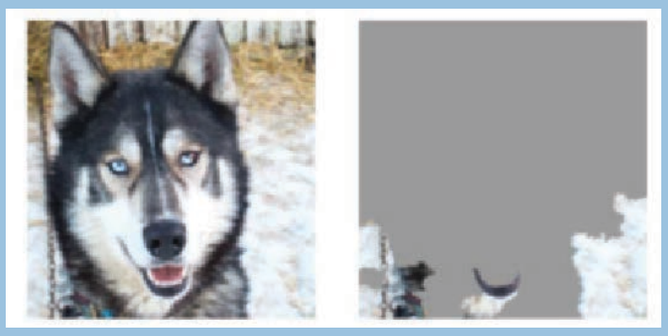
El proyecto surgió de la falta de datos unificados sobre violencia de género en Argentina (la única base de datos oficial abierta es la de la Oficina de violencia doméstica de la Corte Suprema de Justicia y los informes del Registro único de casos de violencia de género, que solo tiene datos hasta 2018). AymurAI puede ayudar a compartir información sobre la violencia de género y cómo se aborda en diferentes juicios.

Actualmente se está implementando AymurAI en el Juzgado Penal 10 de la Ciudad de Buenos Aires. Este Juzgado Penal, liderado por Pablo Casas, promueve, diseña y posibilita la aplicación de políticas de justicia abierta a través de su base de datos pública. Esta base de datos es mantenida por las personas que trabajan en el juzgado. La base de datos tiene alrededor de cinco mil resoluciones legales anónimas fechadas a partir de agosto de 2016, incluidos muchos casos de violencia de género. Contiene 64 categorías con información detallada sobre cada fallo legal, como el tipo de violencia sufrida por la víctima en cada caso, en línea con la Ley N.º 26.485 de Argentina. La base de datos también incluye datos contextuales (por ejemplo, variables

socioeconómicas de las personas involucradas en el conflicto, si el acusado tiene hijos con la víctima y las frases utilizadas durante las agresiones). Los empleados del Juzgado 10 utilizan diferentes herramientas para mantener la base de datos. Por ejemplo, utilizan una herramienta para anonimizar resoluciones legales llamada IA2[186].

Actividad: La IA puede crear riesgos imprevistos que pueden tener resultados potencialmente mortales. Lea el siguiente ejemplo y debata las preguntas con los participantes.

Investigadores de la Universidad de Washington desarrollaron un algoritmo deliberadamente defectuoso que clasificaba las fotos de perros husky y lobos. El algoritmo aprovechó la presencia o ausencia de nieve para distinguir entre huskies domésticos y lobos salvajes. En el conjunto de datos de entrenamiento, los lobos aparecieron en la nieve con más frecuencia que los huskies. Por lo tanto, todas las imágenes de perros lupinos con nieve fueron clasificadas como lobos por el sistema. Como resultado, la IA pudo proporcionar resultados incorrectamente el 50 % de las veces.[187]



Debido a que los píxeles que definen a los lobos son los del fondo nevado (a la derecha), un husky (a la izquierda) se confunde con un lobo. Esta situación es el resultado de una base de aprendizaje inadecuadamente representada.

Fuente: Besse P., Castets-Renard C., Garivier A., Loubes J.-M. (2018). ¿Puede la IA cotidiana ser ética? Machine Learning Algorithm Fairness, disponible en: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3391288

Este ejemplo muestra que podría ser fatal si los sistemas de IA utilizados en campos de alto riesgo se entrenan utilizando una base de aprendizaje representada de manera inadecuada.[188] Por ejemplo, en el sistema de salud los datos de grupos de población específicos tienden a faltar en los datos con los que aprenden las herramientas de aprendizaje automático, lo que significa que la herramienta podría funcionar menos bien para esas comunidades. Para ilustrar esto, un equipo de científicos del Reino Unido descubrió que casi todos los conjuntos de datos de enfermedades oculares provienen de pacientes de América del Norte, Europa y China, lo que significa que es menos probable que los algoritmos de diagnóstico de enfermedades oculares funcionen bien para los grupos raciales de países subrepresentados.[189] Otro ejemplo es que los algoritmos de detección de cáncer de piel tienden a ser menos precisos cuando se usan en pacientes negros porque los modelos de ML están entrenados principalmente en imágenes de pacientes de piel clara.[190]

186 https://www.aymurai.info.
187 Pearson D. (2021). AI biopsy dilemma: Wolf or husky, equity or bias?, disponible en: https://healthexec.com/topics/precision-medicine/ai-biopsy-dilemma-wolf-or-husky-equity-or-bias.
188 Access Now (2018). Human rights in the age of artificial intelligence, disponible en: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf
189 Knight W. (2020). AI Can Help Diagnose Some Illnesses—If Your Country Is Rich, disponible en: https://www.wired.com/story/ai-diagnose-illnesses-country-rich/
190/2018/08/machine-learning-dermatology-skin-color/567619/

Preguntas para el debate:

1. ¿Cuáles fueron los principales factores utilizados por el sistema para diferenciar entre huskies domésticos y lobos salvajes?
2. ¿Hubo algún defecto en este análisis? ¿Por qué?
3. ¿Qué pasaría si los procesos de toma de decisiones de IA implementados en los sistemas de justicia utilizaran algoritmos con fallas similares?

3. ¿Por qué la transparencia algorítmica y la responsabilidad son importantes en el contexto del poder judicial?

La falta de transparencia algorítmica es un tema importante en la vanguardia de los debates sobre IA y derechos humanos. El despliegue de sistemas de IA en el poder judicial está generando preocupaciones sobre cómo evaluar a fondo los efectos a corto y largo plazo, a qué intereses sirven los algoritmos y si son sensibles al contexto para lidiar con el contexto sociocultural en diferentes países.

Esta opacidad de los sistemas de IA es alarmante. Un debate político informado es imposible sin tener la capacidad de comprender cómo funcionan los sistemas de IA. La opacidad en la forma en que los sistemas de IA llegan a sus decisiones y la dificultad para determinar la responsabilidad por sus acciones significan que pueden producirse daños a los derechos humanos cuando se utilizan dichos sistemas.[191]

Al mismo tiempo, también puede darse el caso de que, incluso cuando se puedan explicar las decisiones basadas en IA, los afectados por la decisión puedan no estar de acuerdo con el resultado. En tales situaciones, las partes afectadas deberían tener derecho a un recurso legal. En contraste con los procedimientos sólidos que existen en muchos contextos legales para promover la responsabilidad de las decisiones humanas en el gobierno, desde las leyes de libertad de información hasta las protecciones del debido proceso y los procedimientos de apelación, los algoritmos operan principalmente en una zona libre de responsabilidad. Esta sección analizará la transparencia algorítmica y la rendición de cuentas en el contexto de las operaciones judiciales.

Transparencia algorítmica

Cuando se trata de un sistema de IA, la transparencia se refiere a la cantidad de información que se pone a disposición de la persona usuaria. La estructura del modelo, sus usos previstos, cómo y cuándo se tomaron las decisiones de implementación, quién tomó esas decisiones son parte de la transparencia, que también incluye decisiones de diseño y datos de entrenamiento.[192]

191 Deeks A. (2019). The Judicial Demand for Explainable Artificial Intelligence, 119 Colum. L. Rev. Virginia Public Law and Legal Theory Research Paper No. 2019-51, disponible en: https://ssrn.com/abstract=3440723

192 Malek Md. A. (2021). Transparency in Predictive Algorithms: A Judicial Perspective, disponible en: https://doi.org/10.31124/advance.14699937.v2

Las personas usuarias de un sistema de IA implementado en el poder judicial (por ejemplo, demandantes y acusados) a menudo desconocen cómo se entrenó el sistema de IA y cómo toma decisiones. Por lo tanto, cuando se trata de emprender acciones legales contra los resultados incorrectos y dañinos del sistema de IA, es difícil para los afectados por el uso de sistemas de IA desafiarlos en ausencia de transparencia sobre cómo se diseñó el sistema y cómo funciona.[193]

La necesidad de transparencia algorítmica incluye solicitudes a las empresas para que divulguen sus algoritmos patentados para que puedan revisarse por auditores independientes, reguladores o el público en general antes de la implementación. Sin embargo, es poco probable que se proporcionen los algoritmos o el código de software subyacente al público, ya que las empresas privadas consideran su algoritmo como un activo clave de su propiedad y no están dispuestas a divulgarlo.

El Tribunal de Justicia Europeo ha declarado que las empresas no pueden declarar y argumentar ante los tribunales que no tienen permitido o no pueden divulgar sus algoritmos debido a consideraciones de propiedad intelectual (PI) o secretos comerciales para escapar de su responsabilidad de explicar la IA (en virtud del artículo 22 del RGPD), con la excepción de la IA que sirve a un propósito de seguridad nacional o asuntos penales. Sin embargo, debe tenerse en cuenta que la transparencia adecuada de los sistemas automatizados es complicada y difícil de lograr debido a los frecuentes cambios de algoritmos. Por ejemplo, Google cambia su algoritmo cientos de veces al año.[194] Además, el riesgo de manipular algoritmos aumenta si se hacen públicos.

193 Felzmann H., Fosch-Villaronga E., Lutz C., Tamò-Larrieux A. (2020). Towards Transparency by Design for Artificial Intelligence, Sci Eng Ethics 26, 3333–3361, disponible en: https://doi.org/10.1007/s11948-020-00276-4

194 https://www.aymurai.info.

Estudio de caso: transparencia algorítmica en la práctica

- Reino Unido: La Oficina Central Digital y de Datos del Reino Unido y el Centro de Ética e Innovación de Datos (CDEI) publicaron una de las primeras directrices nacionales de transparencia algorítmica en todo el mundo en 2021. El estándar consiste en una plantilla que se alienta que llenen las organizaciones del sector público para cualquier herramienta algorítmica que involucre directamente al público (como un chatbot) o cumpla con los requisitos específicos basados en el riesgo. La información recopilada sobre las herramientas de IA está disponible en un registro público.[195]

- Francia, los Países Bajos y Nueva Zelanda: los tres países también han desarrollado una guía para ayudar a los funcionarios del sector público a navegar por el uso responsable de los algoritmos. Etalab de Francia apoya a las agencias gubernamentales en la implementación del marco legal para la responsabilidad y la transparencia de los algoritmos del sector público.[196]

- Estados Unidos: varios gobiernos locales en los Estados Unidos han implementado prohibiciones o paradas temporales en el uso de tecnologías algorítmicas, como las tecnologías de reconocimiento facial (FRT), para la aplicación de la ley y la vigilancia. El objetivo principal de estas leyes es abordar las preocupaciones con respecto a la privacidad, pero también hay intersecciones significativas con los problemas de responsabilidad algorítmica. Estas prohibiciones generalmente se establecen a través de la legislación, pero algunas leyes han proporcionado excepciones limitadas a la prohibición, como la información de terceros obtenida a través de FRT. Por ejemplo, un proyecto de ley en San Francisco que prohíbe el uso de FRT solo se aplica a los usos por parte de agencias municipales y excluye el uso por parte de agencias federales, como las de puertos y aeropuertos.[197]

- Chile: GobLab, un laboratorio de innovación dentro de la Facultad de Gobierno de la Universidad Adolfo Ibáñez en Santiago, realizó una extensa investigación sobre el uso de algoritmos por parte del gobierno chileno en colaboración con el Consejo Chileno de Transparencia. Con financiamiento del Banco Interamericano de Desarrollo, el grupo ha redactado y propuesto una regulación que el gobierno está en camino de adoptar luego de las pruebas iniciales de la regulación con varios organismos públicos. La regulación convertirá a Chile en la primera nación de América Latina en adoptar estándares sobre transparencia algorítmica.[198]

- Iniciativas a nivel de ciudad: la transparencia algorítmica en la UE se ha introducido ex ante a nivel local desde octubre de 2020, con las ciudades de Ámsterdam[199], Helsinki[200] y Nantes[201] estableciendo registros que describen los algoritmos empleados en las administraciones de las ciudades. Para garantizar que la IA utilizada por los servicios públicos esté centrada en las personas, los registros indican, entre otras cosas, cómo se procesan los datos, qué peligros existen y si las tecnologías están sujetas a monitoreo humano.[202]

195 Centro de Ética e Innovación de Datos (2023). Algorithmic Transparency Recording Standard Hub. gov.uk, disponible en: https://www.gov.uk/government/collections/algorithmic-transparency-recording-standard-hub
196 Turak H. (2020). Open algorithms: Experiences from France, the Netherlands, and New Zealand. Open Government Partnership, disponible en:https://www.opengovpartnership.org/stories/open-algorithms-experiences-from-france-the-netherlands-and-new-zealand/.
197 Haataja M, van de Fliert L., Rautio P. (2020). Public AI Registers: Realizing AI transparency and civic participation in government use of AI Saidot, disponible en: https://openresearch.amsterdam/en/page/73074/public-ai-registers
198 Aránguiz Villagrán M. (2022). Algorithmic Audit for Decision-Making or Decision Support Systems. Banco Interamericano de Desarrollo, disponible en: http://dx.doi.org/10.18235/0004154
199 Véase: https://algoritmeregister.amsterdam.nl/en/ai-register
200 Véase: https://ai.hel.fi/en/ai-register/
201 Véase: https://data.nantesmetropole.fr/pages/algorithmes_nantes_metropole/
202 Ibid.

La transparencia se complica aún más por el problema de la caja negra de los sistemas de IA (analizado en el Módulo 1). Incluso proporcionar el código fuente del algoritmo puede no ser suficiente. Es necesario explicar cómo se generan los resultados de un algoritmo.[203] Uno de los objetivos regulatorios más importantes para el uso seguro y responsable de los algoritmos dentro del sector público es establecer estándares de explicabilidad.

Caso de estudio: transparencia algorítmica desde la perspectiva de la formulación de políticas públicas: el ejemplo de Francia

En Francia, la Ley para una República Digital de 2016 estipula que siempre que un organismo público someta a los residentes a un procesamiento algorítmico, estos últimos tienen derecho a recibir información sobre: 1) el grado en que el procesamiento algorítmico contribuye a la toma de decisiones; 2) los datos procesados; 3) los parámetros de procesamiento; y 4) las operaciones a las que se aplica dicho procesamiento. La información debe comunicarse a una persona que lo solicite en un lenguaje inteligible y sin infringir los secretos protegidos por la ley.

En 2018, cuando el Consejo Constitucional estaba debatiendo un proyecto de ley para alinear la ley francesa de protección de datos con el RGPD, dictaminó que si un organismo público no puede comunicar los principios operativos de un algoritmo sin poner en peligro los secretos protegidos, no se puede tomar ninguna decisión basada únicamente en dicho algoritmo. Por lo tanto, si una entidad pública basa su decisión únicamente en un algoritmo, no se puede utilizar el secreto comercial para evitar revelar cómo funciona el algoritmo.

Fuente: Décret n° 2017-330 du 14 mars 2017 relatif aux droits des personnes faisant l'objet de décisions individuelles prises sur le fondement d'un traitement algorithmique, disponible en: https://www.legifrance.gouv.fr/jorf/id/JORFTEXT000034194929?r=EILBrO52Ri; véase también: https://www.conseil-constitutionnel.fr/decision/2018/2018765DC.htm .

Responsabilidad algorítmica

La responsabilidad algorítmica se refiere a la capacidad de quienes diseñan, construyen, adquieren o implementan el algoritmo para hacerse responsables de sus acciones e impacto de acuerdo con las políticas y leyes relacionadas con el uso del mismo. Un sistema de gobernanza que responsabiliza a un actor requiere que el mismo pueda explicar y justificar sus decisiones con respecto al algoritmo, y enfrentar las consecuencias si sus acciones son contrarias a la ley.[204]

Responsabilidad por diseño

"Todos los sistemas de IA deben diseñarse para facilitar la capacidad de respuesta y la capacidad de auditoría de extremo a extremo. Esto requiere tanto humanos responsables en el circuito a lo largo de toda la cadena de diseño e implementación, como protocolos de monitoreo de actividades que permitan la supervisión y revisión de extremo a extremo".

Fuente: Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, The Alan Turing Institute, disponible en: https://doi.org/10.5281/zenodo.3240529

203 Ibid.

204 Ada Lovelace Institute, AI Now Institute y Open Government Partnership (2021). Algorithmic Accountability for the Public Sector, disponible en:https://www.opengovpartnership.org/documents/algorithmic-accountability-public-sector/

Los desafíos de responsabilidad algorítmica pueden estar relacionados con el hecho de que el tomador de decisiones (por ejemplo, el juez) no tiene el control de las fuentes de datos (datos obtenidos a través de intermediarios de datos o a través de las autoridades policiales que utilizan herramientas de evaluación de riesgos). Los desafíos también podrían derivarse del hecho de que es muy difícil traducir conceptos algorítmicos complejos (por ejemplo, resultados de algoritmos de agrupación que segmentan poblaciones en función de un gran número de variables de entrada) en conceptos humanos comprensibles (por ejemplo, afiliación racial). Esto podría dar lugar a una interpretación inexacta de los resultados algorítmicos. Los desafíos de responsabilidad algorítmica también pueden desencadenarse por asimetrías de información. Por ejemplo, la opacidad de los algoritmos de ML puede hacer que sea imposible para los interesados conocer y comprender los resultados del proceso de toma de decisiones automatizada (ADM) o incluso ser conscientes de que han sido sometidos a ADM. Además, pueden ocurrir problemas en la etapa de implementación cuando se ingresan datos contradictorios en el sistema para engañarlo y que cometa errores. Consulte las charlas del Módulo 1 sobre temas de ciberseguridad.[205]

4. Enfoque en la identificación biométrica, la tecnología de reconocimiento facial y las falsedades profundas

La adopción de tecnologías de alto riesgo, como el reconocimiento facial y la identificación biométrica, presenta desafíos agravados para los responsables políticos y los reguladores de todo el mundo. Las ONG de derechos humanos también han denunciado la falta de protecciones adecuadas de la privacidad en muchos sistemas nacionales de identidad biométrica, donde se determinó que el acceso a los beneficios sociales y otros servicios gubernamentales estaba supeditado al registro en el sistema.[206]

En este sentido, la Resolución de la Asamblea General de la ONU sobre el derecho a la privacidad en la era digital (2020) se ha referido a *"la piratería y el uso ilegal de tecnologías biométricas"* como *"actos altamente intrusivos que violan el derecho a la privacidad"* que interfieren con la libertad de expresión y opinión, la reunión y asociación pacíficas y la libertad religiosa o de creencias, y *"pueden contradecir los principios de una sociedad democrática, incluso cuando se realizan de manera extraterritorial o a gran escala"*.[207]

205 Parlamento Europeo (2019). A governance framework for algorithmic accountability and transparency, disponible en: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262
206 https://www.ohchr.org/Documents/Issues/Poverty/DigitalTechnology/AmnestyInternational.pdf
207 Asamblea General de la ONU (2020). The right to privacy in the digital age, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/371/75/PDF/N2037175.pdf?OpenElement

Además, un informe de 2021 del Alto Comisionado de las Naciones Unidas para los Derechos Humanos, "El derecho a la privacidad en la era digital", ha pedido una moratoria sobre el uso de tecnologías de reconocimiento facial en espacios públicos, hasta que los gobiernos puedan demostrar que no existen problemas sustanciales relacionados con la precisión o los impactos discriminatorios y que estas tecnologías cumplen con unas normas sólidas de privacidad y protección de datos.[208]

El reconocimiento biométrico se basa en la comparación de la representación digital de una persona de su rostro, huella digital, iris, voz o movimiento con otras representaciones similares almacenadas en una base de datos. Sobre esta base, el sistema decide la probabilidad de que el individuo sea realmente la persona a identificar. Las autoridades de todo el mundo utilizan cada vez más el reconocimiento facial remoto en tiempo real, como una forma de reconocimiento biométrico.[209]

El Alto Comisionado de las Naciones Unidas para los Derechos Humanos ha indicado que "el reconocimiento biométrico en tiempo real plantea serias preocupaciones en virtud del derecho internacional de los derechos humanos".[210] Algunas de estas preocupaciones reflejan problemas con las técnicas predictivas, como la probabilidad de una identificación incorrecta de las personas y los efectos desproporcionados en los miembros de ciertos grupos (la mayoría de las veces marginados).[211] Las personas pueden perfilarse utilizando la tecnología de reconocimiento facial en función de su raza, etnia, origen nacional, género/sexo y otros rasgos.[212]

El reconocimiento biométrico remoto se asocia con una interferencia significativa con el derecho a la privacidad. La información biométrica de una persona es uno de los aspectos clave de su personalidad, ya que expone cualidades distintivas que la diferencian de otras personas.[213] El reconocimiento biométrico remoto permite a las autoridades gubernamentales identificar y rastrear sistemáticamente a las personas en los espacios públicos, y esto puede tener un impacto negativo en el ejercicio de los derechos a la libertad de expresión, reunión pacífica y asociación, y la libertad de movimiento.[214]

208 Consejo de Derechos Humanos de las Naciones Unidas (2021). El derecho a la privacidad en la era digital. The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, disponible en: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

209 Ibid.

210 Consejo de Derechos Humanos de las Naciones Unidas (2020). Impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G20/154/35/PDF/G2015435.pdf?OpenElement

211 Consejo de Derechos Humanos de las Naciones Unidas (2021). El derecho a la privacidad en la era digital. The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, disponible en: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx.

212 Consejo de Derechos Humanos de las Naciones Unidas (2020). Racial discrimination and emerging digital technologies: a human rights analysis, paras. 39–40, disponible en: https://www.ohchr.org/sites/default/files/HRBodies/HRC/RegularSessions/Session44/Documents/A_HRC_44_57_AdvanceEditedVersion.docx

213 Consejo de Derechos Humanos de las Naciones Unidas (2020). Impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests, párr. 33, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G20/154/35/PDF/G2015435.pdf?OpenElement Véase también Tribunal Europeo de Derechos Humanos, Reklos y Davourlis v. Greece, Aplicación N.º 1234/05, Sentencia del 15 de abril de 2009, párr. 40.

214 Véase: Consejo Europeo de Protección de Datos y Supervisor Europeo de Protección de Datos (2021). Dictamen conjunto 5/2021 sobre la propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas sobre la inteligencia artificial (Ley de Inteligencia Artificial), para. 30, disponible en: https://edpb.europa.eu/system/files/2021-06/edpb-edps_joint_opinion_ai_regulation_en.pdf; Consejo de Derechos Humanos de la ONU (2020). Impact of new technologies on the promotion and protection of human rights in the context of assemblies, including peaceful protests, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G20/154/35/PDF/G2015435.pdf?OpenElement; Consejo de Derechos Humanos de la ONU (2019). Vigilancia y derechos humanos, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G19/148/76/PDF/G1914876.pdf?OpenElement

Estudio de casos

El RGPD y los datos biométricos

El RGPD de la UE limita el procesamiento de datos biométricos hasta cierto punto. Solo cuando los datos están conectados a una persona específica, se convierten en datos personales y, por lo tanto, están protegidos por este Reglamento. Según el RGPD, los datos biométricos son "datos personales resultantes de un procesamiento técnico específico relacionado con las características físicas, fisiológicas o de comportamiento de una persona física, que permiten o confirman la identificación única de esa persona física". Por lo tanto, si el reconocimiento biométrico no está dirigido a identificar (sino a categorizar, perfilar o afectar el reconocimiento), es posible que no esté incluido en la definición del RGPD.

De acuerdo con el considerando 51 del RGPD "el tratamiento de fotografías [se considera] datos biométricos solo cuando se trata a través de un medio técnico específico que permite la identificación o autenticación única de una persona física".

Fuente: https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a

El caso de Clearview AI

La Autoridad de Protección de Datos del estado alemán de Hamburgo decidió que Clearview AI procesó ilegalmente los datos biométricos obtenidos y puestos a disposición como servicio. Además, no existía una base legal válida para el procesamiento de datos. El tribunal señaló que Clearview AI ha procesado datos biométricos (en virtud del artículo 4(14) del RGPD), ya que "utiliza un procedimiento matemático especialmente desarrollado para generar un valor hash único del interesado que permite la identificación". El litigio se inició por una queja del interesado, ya que el interesado no había dado su consentimiento para el tratamiento de sus datos biométricos. La Autoridad de Protección de Datos determinó que, aunque Clearview AI no se estableció en la UE, estaba sujeta al RGPD a través del monitoreo de la actividad en línea de los interesados (Artículo 3 (2)(b) GDPR), ya que "no ofrece una instantánea [de individuos], pero evidentemente también archiva fuentes durante un período de tiempo". Se ordenó a Clearview AI que eliminara todos los datos personales del demandante.

Fuente: Future of Privacy Forum (2022). GDPR and the AI Act interplay: Lessons from FPF's ADM Case Law Report, disponible en: https://fpf.org/blog/gdpr-and-the-ai-act-interplay-lessons-from-fpfs-adm-case-law-report

Las tecnologías de reconocimiento facial utilizan imágenes digitales para identificar y validar rostros humanos. Estas tecnologías funcionan identificando rasgos faciales en una imagen de origen y comparándolos en un conjunto de datos. Las tecnologías de reconocimiento facial tienen una amplia gama de usos, aunque se emplean más comúnmente con fines de seguridad, como actividades policiales y de seguridad nacional (por ejemplo, lucha contra el terrorismo). Los avances en IA han mejorado la capacidad y la sofisticación de estas tecnologías en los últimos años, convirtiéndolas en un componente estándar de los bienes de consumo como los teléfonos móviles, lo que permite a las personas usuarias "iniciar sesión" con el rostro.[215]

215 Hill D., O'Connor C. D., Slane A. (2022). Police use of facial recognition technology: The potential for engaging the public through co-constructed policy-making, International Journal of Police Science & Management, 24(3), 325–335, disponible en: https://doi.org/10.1177/14613557221089558

Controversias sobre tecnologías de reconocimiento facial en el sector privado

Varias empresas, incluidas Microsoft e IBM, han sido criticadas por implementar un software de reconocimiento facial que es más preciso para algunos grupos demográficos que para otros. Específicamente, estos sistemas tienden a identificar con precisión a los hombres de piel clara con mucha más frecuencia que a las mujeres de piel más oscura.

Del mismo modo, surgió la controversia cuando el software de etiquetado automático de fotos de Google identificó muchas imágenes de afroamericanos como "gorila" o "mono". Es probable que la causa de estos errores se encuentre en el desarrollo de los modelos algorítmicos. Los modelos fueron entrenados con conjuntos de datos de fotos de personas predominantemente de origen caucásico, y por lo tanto no habían sido entrenados con datos suficientes para identificar a personas no blancas, particularmente mujeres. El trabajo de Joy Buolamwini, científica informática del MIT y fundadora de la Liga de la Justicia Algorítmica, ha llevado a varias empresas a publicar declaraciones que abordan las críticas y reforman sus modelos.

Fuente: https://www.poetofcode.com/

En noviembre de 2021, Meta anunció que estaba "cerrando el sistema de reconocimiento facial en Facebook" citando reglas poco claras de los reguladores. Del mismo modo, IBM dejará de ofrecer su software de reconocimiento facial para ciertas actividades, incluida la vigilancia masiva.

Fuente: Gobierno del Reino Unido (2022). Documento de política que establece un enfoque pro-innovación para regular la IA, disponible en: https://www.gov.uk/government/publications/establishing-a-pro-innovation-approach-to-regulating-ai/establishing-a-pro-innovation-approach-to-regulating-ai-policy-statement

El uso de la tecnología de identificación biométrica y reconocimiento facial en las operaciones judiciales puede convertirse en una caja de Pandora para diferentes tipos de sesgos, como los basados en la raza y el género. El caso de los datos de ImageNet se puede utilizar como ejemplo ilustrativo. Este es un conjunto de datos clave para el desarrollo de aplicaciones de visión artificial, que contiene más del 45 % de las imágenes de EE. UU., en comparación con solo el 3 % de China e India combinadas. Esta falta de diversidad contribuye a las deficiencias de los algoritmos de reconocimiento de imágenes, que interpretan los ojos asiáticos como siempre parpadeantes, etiquetan una imagen de una novia tradicional estadounidense vestida de blanco como "novia", "vestido", "mujer" y "boda", pero etiquetan una imagen de una novia india como "arte de performance" y "disfraz", e identifican erróneamente el género/sexo de las mujeres de piel más oscura con una tasa de error del 35 %, mientras que identifican erróneamente el género/sexo de los hombres de piel más clara con una tasa de error del 0 %.[216]

216 Parlamento Europeo (2019). A governance framework for algorithmic accountability and transparency, disponible en: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262

Si bien la vigilancia masiva habilitada por IA a través del reconocimiento facial implica la recopilación, el almacenamiento y el procesamiento de datos personales (biométricos) (nuestros rostros), también tiene un impacto en nuestra privacidad, identidad y autonomía al abrir la posibilidad de ser observados, rastreados y reconocidos.[217] Las personas pueden sentirse presionadas a adherirse a un estándar en particular debido al efecto "escalofriante" psicológico, que altera el equilibrio de poder entre la persona y el gobierno o la empresa privada que utiliza la tecnología de reconocimiento facial.

Si bien el reconocimiento facial puede tener un efecto más pronunciado en el derecho a la privacidad y la integridad psicológica, se podría argumentar que el seguimiento digital de todos los aspectos de la vida humana (a través de datos de ubicación, datos de IoT de relojes inteligentes, rastreadores de salud, altavoces inteligentes, termostatos, vehículos, etc.) podría tener un impacto similar. La frecuencia cardíaca, la temperatura corporal y otros tipos de reconocimiento biométrico impulsado por la IA miden o incluso pronostican nuestro comportamiento, estado mental y emociones. Esto puede tener un grave impacto en el derecho a la privacidad en el entorno en línea.[218]

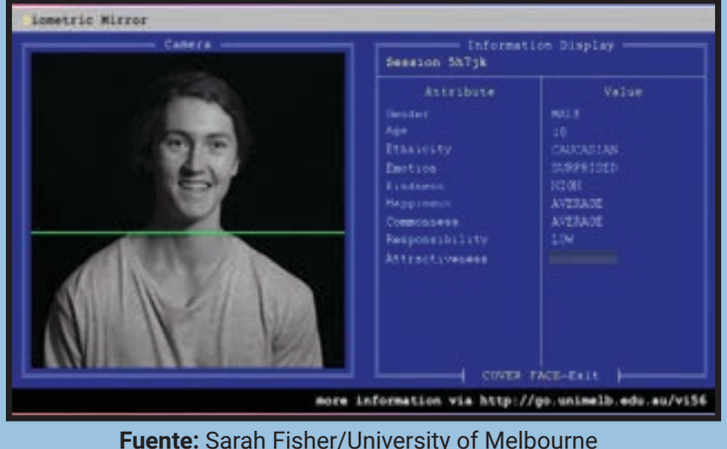
En profundidad: los sistemas de reconocimiento facial pueden identificar erróneamente el género

Los sistemas de IA para "generalizar" a las personas en entornos públicos no son futuristas; ya están en uso en todo el mundo. En Sao Paulo, Brasil, el Instituto Brasileño de Protección al Consumidor (IdeC) cuestionó la instalación y el uso de vallas publicitarias inteligentes que afirman anticipar la emoción, la edad y el género de los pasajeros del metro para proporcionarles "mejores anuncios".[219]

217 Secretaría del CAHAI (2020). Hacia la regulación de los sistemas de IA. Perspectivas globales sobre el desarrollo de un marco legal sobre sistemas de Inteligencia Artificial (IA) basado en los estándares del Consejo de Europa sobre derechos humanos, democracia y Estado de derecho, Estudio del Consejo de Europa, DGI/2020/16, disponible en: https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a.
218 Ibid.
219 Véase: https://idec.org.br/sites/default/files/acp_viaquatro.pdf

Actividad: Los participantes de la capacitación ven el video y analizan las implicaciones sociales de la IA y las tecnologías de reconocimiento facial. También debaten cómo estas tecnologías podrían afectar su trabajo. ¿Cómo afectan las tecnologías de reconocimiento facial a los derechos humanos? ¿Qué grupos son los más vulnerables y susceptibles a las violaciones de los derechos humanos por las tecnologías de reconocimiento facial?

Investigadores con sede en Melbourne pidieron a voluntarios humanos que juzgaran miles de fotos por las mismas características y luego usaron ese conjunto de datos para crear el Espejo biométrico. El Espejo biométrico utiliza la IA para analizar la cara de una persona al escanearla, y luego muestra 14 rasgos sobre ella, como su edad, raza y nivel de atractivo percibido. Utiliza un conjunto de datos abierto de miles de evaluaciones faciales y de crowdsourcing. Sin embargo, este análisis a menudo es falso porque la IA genera el análisis basado en información subjetiva y sesgada proporcionada por voluntarios humanos iniciales.[220]



Fuente: Sarah Fisher/University of Melbourne

Fuente: https://youtu.be/fb_sfhT0mrg

220 Houser K. (2018). Biased AI biometric mirror, disponible en: https://futurism.com/the-byte/biased-ai-biometric-mirror.

Falsedades

Una tecnología de IA particularmente peligrosa que afecta a los derechos humanos son las falsedades (deepfakes). Una falsedad es cualquier forma de medio (video, audio u otro) que ha sido alterado o creado total o parcialmente desde cero.[221] Las máquinas pueden aprender a hacer tareas mirando ejemplos que utilizan redes neuronales. Hay varias tecnologías que se pueden aplicar a esto, pero la más popular se basa en las Redes generativas antagónicas (GAN) y los Modelos de difusión.[222]

221 Van der Sloot B., Wagensveld Y. (2022). Deepfakes: regulatory challenges for the synthetic society. Computer Law & Security Review, disponible en: https://www.sciencedirect.com/science/article/pii/S0267364922000632, disponible en: https://doi.org/10.1016/j.clsr.2022.105716

222 Ibid.

Redes generativas antagónicas (GAN)

Las GAN son un enfoque no supervisado de aprendizaje profundo que puede generar material hiperrealista. Las GAN se utilizan para técnicas de aprendizaje profundo no supervisadas, como la generación de imágenes realistas o conjuntos de datos de imágenes, la realización de traducciones de texto a imagen e imagen a texto, el envejecimiento de las caras y la creación de emojis. Las GAN utilizan dos redes neuronales: un generador que genera nuevas instancias y un discriminador que busca diferenciar estas imágenes falsas, frecuentemente de baja calidad o poco realistas, de la entrada de datos de imágenes reales en el sistema de IA. A través de esta interacción, el generador aprende a producir imágenes cada vez más convincentes y de alta calidad, que finalmente engañan al discriminador para que crea que son parte de los datos de la imagen real.[223]

Modelos de difusión

Los modelos de difusión son modelos generativos que son más avanzados que las GAN en la síntesis de imágenes. Más recientemente, los modelos de difusión se utilizaron en DALL-E 2, el modelo de generación de imágenes de OpenAI y en Imagen de Google.[224] El acceso público a DALL-E se controla a través de una extensa lista de espera y un muro de pago después de varias indicaciones, mientras que Imagen de Google está fuera del alcance del público. La salida de DALL-E se filtra, lo que dificulta la generación de imágenes que contengan violencia, desnudez o rostros realistas.[225]

Sin embargo, el nuevo programa de texto a imagen llamado Stable Diffusion, desarrollado por Stability AI[226], ofrece generación de imágenes de código abierto sin filtro, de uso gratuito para cualquier persona. A continuación se muestra una imagen creada por Stable Diffusion que se creó utilizando el texto exacto "Foto de Bernie Sanders en Mad Max Fury Road (2015), explosiones, cabello blanco, gafas, ropa andrajosa, rasgos faciales simétricos detallados, iluminación dramática".[227]

Imagen: Reddit / Licovoda
Fuente: The Verge (2022). Anyone can use this AI art generator — that's the risk https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data

Como ya se indicó, en 2023, Getty Images presentó una demanda por infracción de derechos de autor contra Stability AI en los Estados Unidos, diciendo que la empresa copió 12 millones de imágenes "sin permiso... o compensación" para entrenar su modelo de IA.

223 AAAS. AAAS, Artificial Intelligence and the Courts: Materials for Judges, disponible en: https://www.aaas.org/ai2/projects/law/judicialpapers
224 O'Connor R. (2022). Introduction to Diffusion Models for Machine Learning, disponible en: https://www.assemblyai.com/blog/diffusion-models-for-machine-learning-introduction/.
225 Véase: https://labs.openai.com/policies/content-policy
226 Véase: https://stability.ai/
227 Vincent J. (2022). Anyone can use this AI art generator — that's the risk https://www.theverge.com/2022/9/15/23340673/ai-image-generation-stable-diffusion-explained-ethics-copyright-data

El verdadero problema asociado a las falsedades es lo sencillo que es generar todo un ecosistema de información falsa. Un video falso, sitios web falsos que alojan el video y generan desinformación y mala información sobre lo que se muestra en el video, cuentas falsas de Twitter que enlazan con el video, cuentas falsas en foros de debates que analizan el contenido del video, cuentas falsas de Instagram que generan memes del video falso. Un entorno de engaño que es multicapa y complejo será extremadamente difícil de penetrar y proporcionar información confiable.[228]

Actividad: los participantes ven el video y analizan cómo las falsedades podrían afectar el trabajo de los operadores judiciales.

Link del video: https://youtu.be/oxXpB9pSETo

Las falsificaciones profundas y todo el ecosistema falsificado que crean ponen en peligro los derechos a un juicio justo, un recurso efectivo y la presunción de inocencia. Podrían usarse como pruebas falsas en los tribunales. Las partes siempre pueden argumentar que la evidencia presentada en su contra es falsa y artificial, y los juicios tomarán más tiempo. Las falsedades también plantean la posibilidad de que un juez acepte erróneamente pruebas fabricadas como confiables.[229] Por lo tanto, el sector judicial debe comenzar a invertir en herramientas digitales que faciliten la evaluación forense de la evidencia de video y audio para determinar que la evidencia no ha sido generada por GAN y Variational AutoEncoders. Por otro lado, la IA tiene el potencial de verificar la autenticidad de la evidencia digital mediante la detección de algoritmos falsos o datos manipulados. El uso de la IA para analizar una imagen o un video podría determinar si se ha manipulado de alguna manera. Sin embargo, esta es todavía un área de investigación en desarrollo.

228 Ibid.
229 Ibid.

5. Actividades

Estas actividades grupales tienen como objetivo alentar a los participantes de la capacitación a discutir diversos desafíos legales y éticos del despliegue de IA en el poder judicial.

Actividad 1

Los participantes en la capacitación leen “State vs. Loomis” en el Módulo 4 y responden a la siguiente pregunta: ¿creen que es apropiado que el tribunal permita que un algoritmo, en el que los actores del sistema legal tienen una visibilidad limitada, desempeñe incluso un papel menor en privar a una persona de su libertad? Por favor, evalúe el impacto ético de esta decisión siguiendo el instrumento de Evaluación de impacto ético de la UNESCO en el Anexo I [céntrese en las partes que se ocupan de la equidad, la no discriminación, la diversidad y la protección de datos y la privacidad].

Actividad 2

Los participantes de la capacitación revisan el material sobre modelos de difusión presentado anteriormente y también leen el siguiente artículo: https://www.theverge.com/2023/2/6/23587393/ai-art-copyright-lawsuit-getty-images-stable-diffusion.

Luego, analizan las implicaciones legales de la demanda presentada por Getty Images contra Stability AI. La demanda se basará en la interpretación de la doctrina de uso justo de los Estados Unidos, que permite el uso no autorizado de obras protegidas por derechos de autor en determinadas circunstancias. La noción de “uso transformador” también puede ser un aspecto significativo. ¿La producción de Stable Diffusion es lo suficientemente distinta de sus datos de entrenamiento? Un estudio reciente ha revelado que el programa memoriza algunas de sus imágenes de entrenamiento y puede repetirlas de manera casi idéntica, aunque en un número relativamente limitado de casos.

Los participantes analizan cómo el desarrollo y la implementación de la IA afectan las normas de derechos de autor en sus propias jurisdicciones.

Actividad 3

Analicen las implicaciones legales y éticas detrás de este caso.

Caso australiano: el alcalde de Victoria se adapta a ChatGPT

El alcalde victoriano, Brian Hood, se está preparando para demandar a OpenAI si no corrige las falsas afirmaciones de ChatGPT de que había cumplido condena en prisión por soborno. Los abogados de Hood enviaron una carta de preocupación a OpenAI el 21 de marzo, dándoles 28 días para corregir los errores, pero OpenAI aún no ha respondido. Las afirmaciones falsas estaban relacionadas con un escándalo de soborno en el extranjero que involucró a una subsidiaria del Banco de la Reserva de Australia a principios de la década de 2000, pero Hood nunca fue acusado de un delito.[230]

230 Byron K. (2023). Victorian mayor readies defamation lawsuit over ChatGPT content, disponible en: https://www.afr.com/technology/vinoctorian-mayor-readies-defamation-lawsuit-over-chatgpt-content-20230405-p5cyh5

Actividad 4

Los participantes de la capacitación leen el texto a continuación sobre cómo las tecnologías de reconocimiento facial pueden invadir el derecho a la privacidad y ven los videos. Luego, analizan cómo se pueden litigar las tecnologías de reconocimiento facial y sus riesgos en virtud de sus leyes nacionales de protección de datos y privacidad.

En mayo de 2020, la Unión Americana de Libertades Civiles (ACLU) presentó una demanda[231] en nombre de organizaciones que representan a víctimas de abuso doméstico, inmigrantes ilegales y trabajadoras sexuales. La organización acusó a Clearview, una empresa de tecnología que desarrolla tecnología de reconocimiento facial, de violar la Ley de Privacidad de Información Biométrica de Illinois (BIPA)[232], un estatuto estatal que impide que las empresas comerciales exploten los identificadores físicos de los ciudadanos, incluido el mapeo computacional de sus rostros, sin consentimiento.[233]

La queja se presentó en el tribunal estatal de Illinois en Chicago después de que el New York Times revelara en enero de 2020 que Clearview estaba desarrollando un sistema de seguimiento y vigilancia basado en un identificador biométrico. La tecnología de reconocimiento facial ha permitido a Clearview adquirir más de tres mil millones de huellas faciales a partir de fotografías web.[234]

Clearview ha proporcionado acceso a esta información a corporaciones privadas, personas adineradas y organizaciones policiales federales, estatales y locales. La empresa afirma que al utilizar esta gran base de datos, puede identificar instantáneamente a las personas con una precisión inigualable, lo que permite una amplia vigilancia clandestina y remota de los estadounidenses.[235]

BIPA exige que las empresas que recopilan, capturan u obtienen un identificador biométrico de un residente de Illinois, como una huella dactilar, una huella facial o un escaneo del iris, primero deben informar a la persona y obtener su consentimiento por escrito. Esto se debe al hecho de que la adquisición forzada de identificadores biométricos inmutables plantea más peligros para la seguridad, privacidad y protección de un individuo que la captura de otros identificadores, como nombres y direcciones. Y registrar la huella facial de una persona, comparable a establecer su perfil de ADN a partir de material genético inevitablemente derramado en una botella de agua, pero distinto de la publicación o transmisión de una fotografía, es un comportamiento, no un discurso, y por lo tanto se rige legítimamente por la ley. Clearview no cumplió con BIPA, privando de derechos de privacidad a varios ciudadanos de Illinois.[236]

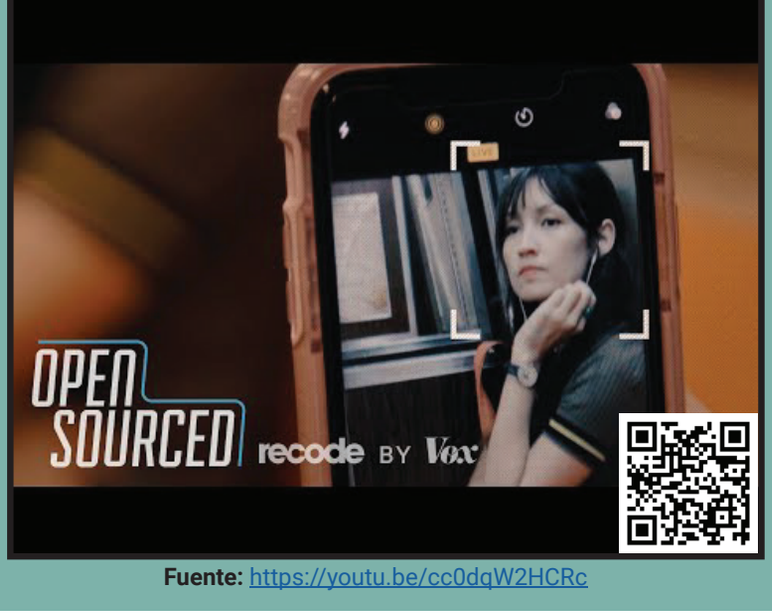
La demanda fue la primera en centrarse en el daño que la tecnología de

231 Alba D. (2020). A.C.L.U. Accuses Clearview AI of Privacy 'Nightmare Scenario', disponible en: https://www.nytimes.com/2020/05/28/technology/clearview-ai-privacy-lawsuit.html.
232 Véase: https://www.aclu-il.org/en/campaigns/biometric-information-privacy-act-bipa
233 Mac R., Hill K. (2022). Clearview AI resuelve la demanda y acepta limitar las ventas de la base de datos de reconocimiento facial. El fabricante de software de reconocimiento facial tiene prohibido en gran medida vender su base de datos de fotos a empresas privadas, disponible en: https://www.nytimes.com/2022/05/09/technology/clearview-ai-suit.html
234 Ibid.
235 Ibid.
236 ACLU (2022). ACLU vs. Clearview AI, disponible en: https://www.aclu.org/cases/aclu-v-clearview-ai

Clearview causaría a los sobrevivientes de abuso doméstico y sexual, inmigrantes indocumentados, comunidades de color y miembros de otras poblaciones vulnerables. Los miembros, clientes y participantes del programa de las organizaciones demandantes han sido expuestos a la grabación facial por Clearview sin su consentimiento y pueden sufrir algunos de los efectos más graves del incomparable programa de monitoreo de Clearview.[237]

El 11 de mayo de 2022, después de que las partes negociaran un acuerdo de conciliación, el tribunal aprobó una orden de consentimiento desestimando este asunto. El elemento fundamental del acuerdo restringe las operaciones de Clearview no solo en Illinois, sino en todo Estados Unidos, prohibiendo permanentemente que Clearview haga accesible su base de datos de huellas faciales a organizaciones privadas. Además, la corporación tiene prohibido durante cinco años vender el acceso a su base de datos a cualquier agencia en Illinois, incluidas las autoridades estatales y municipales.[238]

Fuente: https://youtu.be/s44EFtBoRxY



Fuente: https://youtu.be/cc0dqW2HCRc

237 Ibid.
238 ACLU, EXIBIT 2. signed settlement agreement, disponible en: https://www.aclu.org/legal-document/exhibit-2-signed-settlement-agreement

Actividad 5

Los participantes de la capacitación exploran un caso judicial hipotético que involucra sesgo de IA y responden cómo decidirían el caso si fuera juzgado en su jurisdicción.

Título hipotético del caso: Smith vs. AI Financial Services

Antecedentes: John Smith, un afroamericano, ha presentado una demanda contra AI Financial Services, una importante institución crediticia, alegando prejuicios raciales en el sistema automatizado de aprobación de préstamos de la compañía. Afirma que el sistema de IA negó injustamente su solicitud de hipoteca, lo que le provocó angustia financiera y emocional.

Detalles del caso:

1. **Argumento del Demandante:** John Smith afirma que el sistema de aprobación de préstamos de IA utilizado por AI Financial Services niega desproporcionadamente los préstamos a los afroamericanos, como lo demuestran los datos que muestran una disparidad significativa en las tasas de aprobación de préstamos entre grupos raciales.
2. **Respuesta del Demandado:** AI Financial Services defiende su sistema de IA, afirmando que se basa en criterios financieros objetivos y no considera la raza como un factor en las decisiones de préstamo. Argumenta que cualquier disparidad en las aprobaciones de préstamos se debe a diferencias en los historiales financieros y la solvencia de los solicitantes.

Examen del sistema de IA: durante el juicio, ambas partes traen testigos expertos para examinar el sistema de IA:

1. **Experto del Demandante:** un experto en ética de IA testifica que los datos de capacitación del sistema de IA tenían un sesgo racial inherente, que influyó en su toma de decisiones. Presenta evidencia de casos similares en los que los sistemas de IA han exhibido un comportamiento discriminatorio.
2. **Experto del Demandado:** el experto en IA del demandado argumenta que el sistema de IA fue diseñado para ser neutral en cuanto a la raza y que cualquier sesgo en los datos de entrenamiento no fue intencional. Destaca los rigurosos procesos de prueba y validación a los que se sometió la IA antes de su implementación.

Función del tribunal: el juez debe determinar si el sesgo de IA desempeñó un papel en la denegación del préstamo de John Smith y, de ser así, si AI Financial Services es responsable de discriminación. Entre las consideraciones clave se incluyen:

1. **Transparencia del sistema de IA:** el tribunal evalúa la transparencia del proceso de toma de decisiones del sistema de IA y si el demandado reveló adecuadamente su uso de IA a los solicitantes de préstamos.

2. **Intención vs. Impacto:** el juez distingue entre la discriminación intencional y el impacto dispar resultante del sesgo de la IA, que aún puede ser ilegal según las leyes contra la discriminación.
3. **Esfuerzos de mitigación:** el tribunal examina si AI Financial Services tomó medidas razonables para mitigar el sesgo en su sistema de IA y si abordó rápidamente cualquier problema identificado.

Resultado: el tribunal falla a favor de John Smith, dictaminando que el sistema de IA utilizado por AI Financial Services exhibió un sesgo que resultó en un impacto dispar en los solicitantes afroamericanos. La sentencia incluye una compensación financiera para John Smith y una orden judicial que requiere que AI Financial Services revise y modifique sus algoritmos de IA para garantizar el cumplimiento de las leyes contra la discriminación.

Este caso hipotético destaca los complejos problemas legales que rodean el sesgo de la IA en los préstamos y la importancia de la transparencia, la equidad y la responsabilidad en el uso de los sistemas de IA, especialmente cuando afectan los derechos de las personas y el acceso a los servicios financieros.

6. Recursos

1. Alang N. (2017). Turns Out Algorithms are Racist, disponible en: https://newrepublic.com/article/144644/turns-algorithms-racist/
2. Angwin J., Larson J., Mattu S., Kirchner L. (2016). Machine bias,, disponible en: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Buolamwini J., Gebru T. (2018). Sombras de género: disparidades interseccionales de precisión en la clasificación comercial de género. Actas de la 1.ª Conferencia sobre Equidad, Rendición de Cuentas y Transparencia, PMLR, 81, 77–91, disponibles en: https://proceedings.mlr.press/v81/buolamwini18a.html
4. Commission Nationale de l'Informatique et des Libertés (2022). Asking the right questions before using an artificial intelligence system, disponible en: https://www.cnil.fr/en/asking-right-questions-using-artificial-intelligence-system
5. Edwards L., Veale M. (2017). ¿Esclavo del algoritmo? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For, 16 Duke Law & Technology Review, 18, disponible en: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2972855
6. Servicio de Investigación del Parlamento Europeo (2019). A governance framework for algorithmic accountability and transparency, disponible en: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624262/EPRS_STU(2019)624262_EN.pdf
7. Green B., Chen Y. (2019). Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments, Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 90–99, disponible en: https://doi.org/10.1145/3287560.3287563
8. Hart R. (2017). If you're not a white male, artificial intelligence's use in healthcare could be dangerous, disponible en: https://qz.com/1023448/if-youre-not-a-white-male-artificial-intelligences-use-in-healthcare-could-be-dangerous
9. Kleinberg J. , Lakkaraju H., Leskovec J., Ludwig J., Mullainathan S. (2017). Human Decisions and Machine Predictions, disponible en: https://www.cs.cornell.edu/home/kleinber/w23180.pdf
10. Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector, The Alan Turing Institute, disponible en: https://doi.org/10.5281/zenodo.3240529
11. UTS Human Technology Institute report (2022). outlining a Model Law for facial recognition: https://www.uts.edu.au/human-technology-institute/projects/facial-recognition-technology-towards-model-law
12. Whittlestone J., Nyrup R., Alexandrova A., Cave S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions, Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), Association for Computing Machinery, 195–200, disponible en: https://doi.org/10.1145/3306618.3314289

Módulo 4

Derechos humanos e IA

El módulo cuatro proporcionará un análisis en profundidad de algunos derechos humanos afectados por la IA, como (i) el derecho al acceso a los tribunales, a un juicio justo y el debido proceso, (ii) un recurso efectivo, (iii) los derechos a la protección contra la discriminación, (iv) la libertad de expresión, (v) el derecho a la privacidad y la protección de datos, y (vi) el acceso a la información. El módulo también ofrece una visión general de los principales enfoques de gobernanza de la IA: basada en el riesgo y basada en los derechos humanos.

¿Qué va a aprender?

Después de completar este módulo, los participantes podrán:

- Comprender y explicar los casos de posibles violaciones de los derechos humanos mediante el uso de ADM e IA: (i) el derecho al acceso a los tribunales, a un juicio justo y al debido proceso, (ii) un recurso efectivo, (iii) el derecho a la protección contra la discriminación, (iv) la libertad de expresión, (v) el derecho a la privacidad y la protección de datos y (vi) el acceso a la información.
- Comprender los enfoques clave de gobernanza de la IA: basados en el riesgo y en los derechos humanos.

1. Introducción a los derechos humanos y la IA

Existe una fuerte correlación entre la democracia, el Estado de derecho y los derechos humanos. Las instituciones democráticas sólidas y responsables, los procesos de toma de decisiones inclusivos y transparentes y un poder judicial independiente e imparcial que defienda el Estado de derecho son requisitos previos para defender los derechos humanos.

Los derechos humanos son las libertades y derechos fundamentales que toda persona tiene desde su nacimiento hasta su muerte. Los derechos humanos sostienen y defienden la dignidad inalienable de cada persona independientemente de su raza, etnia, género, edad, orientación sexual, clase, religión, nivel de discapacidad, idioma, nacionalidad o cualquier otro atributo. Los gobiernos están obligados a salvaguardar, defender y cumplir los derechos humanos. Las personas tienen derecho a recursos legales que prevean la reparación de cualquier violación de los derechos humanos.

La Carta Internacional de Derechos[239] representa un cuerpo de derecho internacional de los derechos humanos que incluye nueve tratados importantes de derechos humanos; instrumentos regionales de derechos humanos en América, África y Europa; se ha incorporado a las constituciones nacionales y las leyes nacionales; y la jurisprudencia y el derecho consuetudinario.[240]

Los instrumentos intergubernamentales no vinculantes, como los Principios Rectores de las Naciones Unidas sobre las empresas y los derechos humanos, también[241] han abordado la cuestión de la responsabilidad de las partes interesadas del sector privado en el contexto de los derechos humanos.

Los derechos humanos ofrecen un conjunto de estándares básicos globales basados en principios como la igualdad, la autonomía y la dignidad humana. Estos principios y el marco jurídico que los acompaña imponen obligaciones jurídicamente vinculantes a las naciones de respetar, salvaguardar y defender los derechos humanos.[242]

El derecho internacional de los derechos humanos exige que los estados nacionales proporcionen un recurso efectivo cuando una persona sufre una violación de sus derechos humanos. Los recursos efectivos comprenden recursos judiciales y administrativos, como ordenar una indemnización o una disculpa, y medidas preventivas que pueden incluir cambios en la ley, la política y la práctica. Las obligaciones en materia de derechos humanos también requieren que los Estados establezcan mecanismos eficaces para evitar que se vulneren los derechos humanos.

239 Compuesto por la Declaración Universal de los Derechos Humanos, el Pacto Internacional de Derechos Civiles y Políticos y el Pacto Internacional de Derechos Económicos, Sociales y Culturales.

240 Baluarte D. C., De Vos C. M. (2010). From Judgment to Justice: Implementing International and Regional Human Rights Decisions, Open Society Justice Initiative, Open Society Foundations: New York, disponible en: https://www.justiceinitiative.org/uploads/62da1d98-699f-407e-86ac-75294725a539/from-judgment-to-justice-20101122.pdf

241 Consejo de Derechos Humanos de las Naciones Unidas (2011). Guiding Principles on Business and Human Rights, disponible en: https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf

242 El artículo 2(3) del Pacto Internacional de Derechos Civiles y Políticos requiere que cada Estado parte garantice que una persona cuyos derechos del Pacto hayan sido violados tenga un recurso efectivo, y que este recurso se haga cumplir. Véase también: Comité de Derechos Humanos de la ONU (2004). The Nature of the General Legal Obligation Imposed on States Parties to the Covenant, disponible en: https://www.refworld.org/docid/478b26ae2.html

El marco del derecho internacional de los derechos humanos es un medio establecido para garantizar la protección de los derechos en general y en el entorno digital, incluidos los derechos a la igualdad y la no discriminación. Su naturaleza como un conjunto de estándares procesables se presta especialmente bien a las tecnologías que trascienden las fronteras nacionales, como la IA. Un enfoque basado en los derechos humanos proporciona orientación normativa a los desarrolladores de IA para defender la dignidad humana, independientemente de la jurisdicción.

Las leyes de derechos humanos pueden informar el desarrollo de salvaguardias técnicas y políticas en el despliegue de IA. En este sentido, en 2019, el Consejo de Derechos Humanos (CDH) aprobó la primera resolución (41/11) sobre "Nuevas y emergentes tecnologías digitales y derechos humanos".[243] La resolución reconoce la necesidad de abordar mejor todo el espectro de las implicaciones de las nuevas tecnologías en los derechos humanos para seguir siendo relevantes en la era digital.

En 2021, el Consejo adoptó la resolución 47/23, enfatizando la importancia de un enfoque basado en los derechos humanos para desarrollar y desplegar tecnologías digitales innovadoras. La resolución señala que las nuevas tecnologías tienen el potencial de ofrecer múltiples oportunidades para promover los derechos humanos al contribuir positivamente a la construcción de instituciones democráticas y la resiliencia de la sociedad civil, así como al logro de los Objetivos de Desarrollo Sostenible (ODS). Los defensores de los derechos humanos y los desarrolladores de tecnología, así como los gobiernos, deben seguir siendo ágiles para abordar las preocupaciones de derechos humanos planteadas por la IA, utilizando protecciones e instrumentos basados en las normas y marcos de derechos humanos existentes.[244]

Para que la IA beneficie al bien público, su diseño e implementación debe, como mínimo, evitar dañar los valores humanos fundamentales garantizados por el derecho internacional de los derechos humanos, que proporciona un marco sólido para la protección de estos valores. La IA, si se implementan salvaguardias adecuadas, también podría ser un facilitador clave para mejorar y promover los derechos humanos.

¿Cómo puede la IA ayudar en la protección y el cumplimiento de los derechos humanos?

Los sistemas de IA tienen numerosas aplicaciones que pueden ayudar en el cumplimiento de los derechos humanos. Por ejemplo, los sistemas de IA se utilizan para analizar patrones de escasez de alimentos para combatir el hambre, mejorar el diagnóstico y el tratamiento médico o hacer que los servicios de salud sean más accesibles.

243 ONU, Consejo de Derechos Humanos (2019). New and emerging digital technologies and human rights, disponible en: https://digitallibrary.un.org/record/3834165

244 DiPLO (2022). Promoting and Protecting Human Rights in the Digital Era, disponible en: https://www.diplomacy.edu/event/promoting-and-protecting-human-rights-in-the-digital-era/

El Módulo 2 brindó una descripción general de cómo la IA puede ayudar a los operadores judiciales a través del descubrimiento electrónico y la revisión de documentos, el análisis predictivo y el soporte de ADM, las herramientas de evaluación de riesgos, la resolución de disputas, la IA generativa, el reconocimiento y análisis de idiomas y la gestión de casos y archivos digitales. El poder judicial, como institución pública, está sujeto a un estándar más alto cuando se trata del comportamiento de los operadores judiciales, y los jueces en particular, hacia los individuos y la sociedad. Esto se ha reflejado en los principios del Estado de derecho, como la justificación, la proporcionalidad y la igualdad. Por un lado, la IA puede aumentar la eficiencia de los operadores judiciales, por otro lado, también puede erosionar la legitimidad procesal y la confianza en las instituciones democráticas y la autoridad de la ley.

Sin las barandas adecuadas, la IA también podría invadir los derechos humanos

Por ejemplo, el sesgo no detectado podría estar presente en los algoritmos de ML que predicen la reincidencia. O el despliegue de IA podría utilizarse para limitar la libertad de expresión de las personas o su capacidad para participar en actividades políticas o para identificar a los disidentes políticos.[245] La IA también podría dañar los derechos humanos en situaciones en las que se utilizan datos de capacitación de baja calidad, diseño de sistemas o interacciones complejas entre el sistema de IA y su entorno. Un ejemplo de ello es la exacerbación algorítmica del discurso de odio o la incitación a la violencia en línea. Otro ejemplo es la amplificación de la desinformación y la mala información, que podría afectar el derecho a participar en los asuntos políticos y públicos, especialmente durante las elecciones. La escala y el impacto probables del daño estarán vinculados a la escala y el impacto potenciales de las decisiones de cualquier sistema de IA específico. Al mismo tiempo, es importante tener en cuenta que la IA se puede utilizar para identificar el discurso de odio y ayudar a eliminar el contenido relacionado con la promoción del terrorismo.

Numerosas aplicaciones de la IA tienen el potencial de afectar directamente la igualdad de acceso a los derechos fundamentales, incluido el derecho a la privacidad y la protección de la información personal, el derecho al acceso a la justicia y el derecho a un juicio justo, particularmente en lo que respecta a la presunción de inocencia y la carga de la prueba, el derecho al empleo, la educación, la vivienda y la salud, así como el derecho a los servicios públicos y el bienestar. Si no van acompañadas de las salvaguardias adecuadas contra los prejuicios, las tecnologías de la IA podrían contribuir a denegar de forma desproporcionada el acceso a los derechos a las mujeres, las minorías y los que ya son los más vulnerables y marginados.[246]

245 Asamblea General de la ONU (2018). Promotion and protection of the right to freedom of opinion and expression. Nota del Secretario General, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N18/270/42/PDF/N1827042.pdf?OpenElement

246 Consejo de Europa (2019). Preventing discrimination caused by the use of artificial intelligence, disponible en: https://pace.coe.int/en/files/28809.

Por ejemplo, el uso de sistemas biométricos o de reconocimiento facial en espacios públicos podría permitir la vigilancia masiva que invade los derechos humanos.[247] Según el informe 'Regulating Biometrics' (2020) del AI Now Institute, la tecnología de reconocimiento[248] facial no es un sustituto adecuado de las huellas dactilares. Las tecnologías de reconocimiento facial muestran resultados de bajo rendimiento y altas tasas de error para "mujeres negras, minorías de género, jóvenes y ancianos, miembros de la comunidad con discapacidad y trabajadores manuales".[249]

A menudo, el despliegue de IA por parte de los organismos encargados de hacer cumplir la ley puede invadir el debido proceso y la igualdad de derechos de protección. Por ejemplo, si el sistema de IA se utiliza para pruebas de ADN que involucran el procesamiento de datos de salud confidenciales y evaluaciones de riesgos de la justicia penal que podrían estar sesgadas hacia ciertas poblaciones en función del género/sexo, raza, etnia, etc.

¡Recordatorio!

Como hemos visto, las herramientas de vigilancia predictiva o reconocimiento facial no pueden ser una predeterminación de culpabilidad o pruebas suficientes para refutar la presunción de inocencia. Una predicción estadística no puede ser una causa de arresto o, según el derecho consuetudinario, una sospecha razonable, o un paso más allá, una causa probable, y está lejos de ser un caso prima facie, y mucho menos una evidencia inculpatoria. Su valor de inteligencia no puede exceder el otorgado a la información policial o de inteligencia y, por lo tanto, no tendría valor probatorio. Utilizarlo como única fuente violaría el principio de presunción de inocencia.

El uso de la IA debe estar dirigido hacia el principio de beneficencia o hacer el bien, el mejoramiento y el progreso de la humanidad. Por lo tanto, el desarrollo y el uso de los sistemas de IA deben estar dirigidos al beneficio y bienestar de la sociedad y la civilización humana para la mejora de las condiciones de vida, la salud, el trabajo y el desarrollo de las capacidades físicas y mentales.

Aunque se puede esperar que la estructura básica y el marco institucional para la protección de los derechos humanos, que están bien establecidos y son reconocidos universalmente, desarrollen respuestas efectivas a muchas de las amenazas y desafíos provocados por el creciente poder de la automatización digital y la inteligencia artificial, hay varias razones por las que los mecanismos existentes de aplicación de los derechos humanos pueden requerir una revitalización para proporcionar una protección efectiva: en primer lugar, muchos de los derechos son difíciles de afirmar en la práctica, debido a la opacidad de muchos de los sistemas sociotécnicos en los que están integradas estas tecnologías. En segundo lugar, nuestra comprensión del alcance y el contenido de los derechos existentes se

247 Human Rights Watch (2020). Argentina: Child Suspects' Private Data Published Online, disponible en: https://www.hrw.org / news/2020/10/09/argentina-child-suspects-private-data-published-online
248 Kak A. (2020). Regulación de la biometría. Global Approaches and Urgent Questions, disponible en: https://ainowinstitute.org/publication/regulating-biometrics-global-approaches-and-open-questions
249 Ibid.

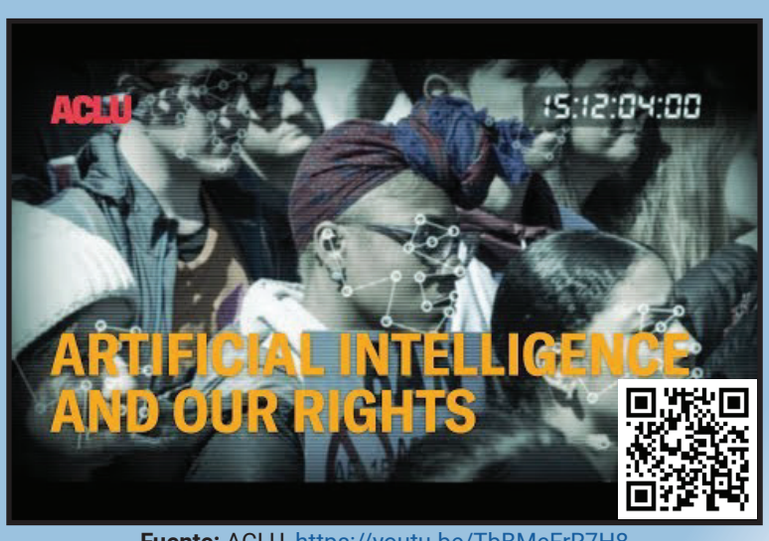
desarrolló en una era previa a la creación de la red. Así concebidos, estos derechos podrían no proporcionar una protección integral contra toda la gama de amenazas y riesgos a los que pueden dar lugar estas tecnologías, particularmente en relación con la discriminación y los intentos ilegítimos de engañar y manipular a las personas utilizando "tecnologías persuasivas"[250].

La Figura 13 a continuación ofrece una descripción general de algunos derechos humanos cubiertos en este Kit de herramientas que podrían verse afectados por el despliegue de IA en general.

Figura 13. Derechos humanos afectados por la IA

Fuente: OECD, https://www.oecd-ilibrary.org/sites/ba682899-en/

Actividad: los participantes de la capacitación ven el video y debaten cómo la IA puede afectar los derechos humanos.



Fuente: ACLU, https://youtu.be/TbBMeFrR7H8

250 La tecnología persuasiva es una "tecnología creada específicamente para cambiar las opiniones, actitudes o comportamientos de sus usuarios para cumplir sus objetivos", véase: Centre for Humane technology (2021). Tecnología persuasiva. How does technology use design to influence my behavior?, disponible en: https://assets.website-files.com/5f0e1294f002b15080e1f2ff/612f8e3e010ff2e211c92019_2%20-%20Persuasive%20Technology%20Issue%20Guide.pdf

Ventajas del enfoque de derechos humanos para el desarrollo y la implementación de IA

Los mecanismos institucionales del derecho de los derechos humanos proporcionan la dirección y la base para garantizar el desarrollo y el uso ético y centrado en el ser humano de la IA en la sociedad. Los operadores judiciales pueden recomendar la debida diligencia en materia de derechos humanos, como las evaluaciones de impacto en los derechos humanos (HRIA), para evaluar los riesgos que plantea el despliegue de IA en los derechos humanos. Cuanto mayor sea el riesgo para los derechos humanos, en mayor medida la IA podría considerarse no apta para su uso.

Las evaluaciones de impacto en los derechos humanos pueden ayudar a identificar grupos o comunidades vulnerables o en riesgo en relación con la IA. Algunas personas o comunidades pueden estar subrepresentadas debido, por ejemplo, al uso limitado de teléfonos inteligentes y a la ausencia de sus datos en los conjuntos de datos utilizados para capacitar a los sistemas de IA. El enfoque basado en los derechos humanos puede proporcionar reparación a aquellos cuyos derechos son violados. Ejemplos de remedios incluyen el cese de la actividad, el desarrollo de nuevos procesos o políticas, una disculpa o una compensación monetaria.

Existen cinco ventajas clave en el aprovechamiento de los marcos de derechos humanos en el contexto de la IA.[251]

- Con el tiempo, se ha desarrollado una vasta infraestructura internacional, regional y nacional de derechos humanos, y existen instituciones establecidas que pueden ayudar a la realización de los derechos humanos en el contexto de la inteligencia artificial. Esta infraestructura incluye organizaciones intergubernamentales, tribunales, organizaciones no gubernamentales, instituciones académicas y otras instituciones y comunidades donde se pueden hacer valer los derechos humanos y buscar reparación.
- Un cuerpo integral de derecho nacional, regional e internacional ha operacionalizado la aplicación de los derechos humanos en el ámbito digital.
- Los derechos humanos dan un lenguaje universal para cuestiones que trascienden las fronteras nacionales, como la IA. Junto con la infraestructura de derechos humanos, esto puede ayudar a llegar e incluir a una gama más amplia de partes interesadas.
- Los derechos humanos gozan de legitimidad y apoyo generalizados en todo el mundo. La mera percepción de que un actor puede violar los derechos humanos podría ser significativa debido a los costos sustanciales de reputación asociados con tal percepción
- Muchos estados tienen algún tipo de marco de derechos humanos, incluso si no tienen un marco de protección de datos; por lo tanto, utilizar el marco de derechos humanos como base haría que el proceso fuera más inclusivo.

251 Véase: https://www.oecd-ilibrary.org/sites/969ff07f-en/index.html?itemId=/content/component/969ff07f-en

Un desafío relacionado con el enfoque de derechos humanos para el desarrollo y la implementación de IA es el hecho de que su aplicación está vinculada a las jurisdicciones. Los Demandantes a menudo deben demostrar su capacidad legal en una jurisdicción en particular. Cuando los problemas involucran a grandes corporaciones internacionales y sistemas de IA que abarcan numerosas jurisdicciones, estos enfoques pueden no ser óptimos.[252]

Tabla 5. Instrumentos internacionales clave relacionados con el derecho a la privacidad en general, y en el entorno en línea en particular.

Tratados	
1	Declaración Universal de Derechos Humanos[253]
2	Pacto Internacional de Derechos Civiles y Políticos[254]
3	Convención Internacional sobre la Eliminación de todas las formas de discriminación racial[255]
4	Directrices de la OCDE que regulan la protección de la privacidad y los flujos transfronterizos de datos personales[256]
5	Convenio del Consejo de Europa para la protección de las personas con respecto al tratamiento automatizado de datos de carácter personal (Convenio 108 / Convenio 108+)[257]
Normas	
6	Las Directrices de las Naciones Unidas relativas a los archivos informatizados de datos personales (ONU, 1990)[258]
7	Las Normas Internacionales de Privacidad y Protección de Datos (la Resolución de Madrid)[259]
8	La Recomendación de la OCDE sobre la Gestión de riesgos de seguridad digital para la prosperidad económica y social[260]
9	Directrices de la OCDE que regulan la protección de la privacidad y los flujos transfronterizos de datos personales[261]
10	Los Principios de la ONU sobre Protección de Datos Personales y Privacidad (2018)[262]
11	La Resolución de la Asamblea General de la ONU sobre el Derecho a la Privacidad en la Era Digital de 2014[263]
Otros documentos	
12	Informe del Relator Especial sobre la promoción y la protección de los derechos humanos y las libertades fundamentales en la lucha contra el terrorismo de 2014[264]
13	Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión de 2018[265]
14	La Resolución de la Asamblea General de la ONU sobre el derecho a la privacidad en la era digital (2020) se ha referido a *"la piratería y el uso ilegal de tecnologías biométricas"* como *"actos altamente intrusivos que violan el derecho a la privacidad"* que interfieren con la libertad de expresión y opinión, la reunión y asociación pacíficas y la libertad religiosa o de creencias, y *"pueden contradecir los principios de una sociedad democrática, incluso cuando se realizan de manera extraterritorial o a gran escala"*.[266]
15	Un informe de 2021 del Alto Comisionado de las Naciones Unidas para los Derechos Humanos, "El derecho a la privacidad en la era digital", ha pedido una moratoria sobre la implementación de tecnologías de reconocimiento facial en espacios públicos, hasta que los gobiernos puedan demostrar que no existen problemas sustanciales relacionados con la precisión o los impactos discriminatorios y que estas tecnologías cumplen con unas normas sólidas de privacidad y protección de datos.[267]
16	UNSDG Privacidad, ética y protección de datos: nota orientativa sobre Big Data para la consecución de la Agenda 2030 (2017)[268]
17	Compendio de las Naciones Unidas de políticas de privacidad y protección de datos[269]

252 Ibid.
253 Véase: https://www.un.org/en/about-us/universal-declaration-of-human-rights
254 Oficina de Derechos Humanos de la ONU (1976). Pacto Internacional de Derechos Civiles y Políticos, disponible en: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights
255 Oficina de Derechos Humanos de la ONU (1965). Convención internacional sobre la eliminación de todas las formas de discriminación racial, disponible en: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-convention-elimination-all-forms-racial
256 OCDE (2002). OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, disponible en: https://www.oecd-ilibrary.org/docserver/9789264196391-en.pdf?expires=1695655643&id=id&accname=ocid195767&checksum=923738DCA1AEE95B3D260E41902AC30D
257 Consejo de Europa (CdE) (2018). Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108+), disponible en: https://www.coe.int/en/web/data-protection/convention108-and-protocols
258 Joinet L. (1988). Directrices para la regulación de los archivos informatizados de datos personales: informe final, disponible en: https://digitallibrary.un.org/record/43365?ln=en
259 Véase: https://www.dataguidance.com/opinion/international-madrid-resolution
260 OCDE (2015). Digital Security Risk Management for Economic and Social Prosperity: OECD Recommendation and Companion Document, disponible en:https://www.oecd.org/publications/digital-security-risk-management-for-economic-and-social-prosperity-9789264245471-en.htm
261 Véase: https://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm
262 Véase: https://unsceb.org/privacy-principles
263 Consejo de Derechos Humanos de las Naciones Unidas (2014). The right to privacy in the digital age: report of the Office of the United Nations High Commissioner for Human Rights, disponible en: https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F %2Fwww.ohchr.org%2Fsites%2Fdefault%2Ffiles%2FDocuments%2FIssues%2FDigitalAge%2FA-HRC-27-37_en.doc&wdOrigin=BROWSELINK
264 Véase: https://www.ohchr.org/en/special-procedures/sr-terrorism
265 Asamblea General de la ONU (2018). Promoción y protección del derecho a la libertad de opinión y expresión, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N18/270/42/PDF/N1827042.pdf?OpenElement
266 Asamblea General de la ONU (2020). The right to privacy in the digital age, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N20/371/75/PDF/N2037175.pdf?OpenElement
267 Consejo de Derechos Humanos de las Naciones Unidas (2021). El derecho a la privacidad en la era digital. The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, disponible en: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx
268 Véase: https://unsdg.un.org/resources/data-privacy-ethics-and-protection-guidance-note-big-data-achievement-2030-agenda
269 PNUD (2021). COMPENDIO DE POLÍTICAS DE PRIVACIDAD Y PROTECCIÓN DE DATOS Y OTRAS ORIENTACIONES RELACIONADAS DENTRO DE LA ORGANIZACIÓN DE LAS NACIONES UNIDAS Y OTROS ÓRGANOS SELECCIONADOS DEL COMMUNIT INTERNACIONAL, disponible en: https://unstats.un.org/legal-identity-agenda/documents/Paper/data_protecton_%20and_privacy.pdf

Derecho al acceso a los tribunales, a un juicio justo y al debido proceso

"Todas las personas son iguales ante los tribunales y cortes de justicia. Toda persona tendrá derecho a ser oída públicamente y con las debidas garantías por un tribunal competente, independiente e imparcial, establecido por la ley, en la substanciación de cualquier acusación de carácter penal formulada contra ella [...] Toda persona acusada de un delito penal tendrá derecho a que se presuma su inocencia hasta que se demuestre su culpabilidad de acuerdo con la ley."

– Artículo 14 del Pacto Internacional de Derechos Civiles y Políticos

Cuando se trata de la aplicación de la ley y el sistema legal, el potencial de la IA para reforzar o amplificar los sesgos existentes es una preocupación importante. Los derechos a la libertad, la seguridad y un juicio justo pueden ser infringidos cuando la libertad física o la seguridad personal de un individuo está en juego, como con la vigilancia predictiva, la evaluación del riesgo de reincidencia y la sentencia. Como ya se debatió, los sistemas de IA de «caja negra» hacen que sea imposible para los profesionales legales, como jueces, abogados y fiscales, comprender la justificación detrás de los resultados del sistema, lo que complica la justificación y la apelación de la decisión.[270]

La IA y la toma de decisiones automatizada (ADM) tienen un impacto sustancial en la vida de las personas, y con frecuencia pueden restringir el derecho a participar, impugnar o cuestionar el resultado de la decisión o sus aportes. A menudo, los sistemas de IA, debido a su naturaleza de «caja negra», no pueden producir una explicación inteligible y comprensible para los humanos de sus decisiones. Estos sistemas también pueden tener sesgos integrados que limitan el acceso de los datos invisibles y de los grupos marginados a los tribunales y la justicia.

Las herramientas para la evaluación de riesgos penales, por ejemplo, se ofrecen como instrumentos para ayudar a los jueces en las decisiones de sentencia. Aunque las autoridades atribuyen un nivel de culpa potencial al clasificar a una persona como de alto o bajo riesgo de reincidencia, esto podría estar en desacuerdo con el derecho a un jurado imparcial y la presunción de inocencia.

270 Secretaría del CAHAI (2020). Hacia la regulación de los sistemas de IA. Perspectivas globales sobre el desarrollo de un marco legal sobre sistemas de inteligencia artificial (IA) basado en los estándares del Consejo de Europa sobre derechos humanos, democracia y Estado de derecho, Estudio del Consejo de Europa, DGI/2020/16, disponible en: https://rm.coe.int/prems-107320-gbr-2018-compli-cahai-couv-texte-a4-bat-web/1680a0c17a.

271 AccessNow (2018). AI and human rights, disponible en: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

El software de vigilancia predictiva también puede reflejar sesgos sociales y puede suponer el riesgo de utilizar datos históricos para introducir sesgos y atribuir falsamente la culpa.[271] Hay varios casos documentados en los que el uso de algoritmos de IA en la vigilancia predictiva, la evaluación de riesgos y las sentencias ha dado lugar a resultados subóptimos en el sistema de justicia penal. En muchos casos, el uso de IA para la puntuación de riesgo de los acusados y los esfuerzos de vigilancia predictiva se anuncian como intentos bien intencionados de eliminar el posible sesgo humano de los jueces en sus decisiones de sentencia y fianza, al tiempo que se asignan recursos policiales limitados para prevenir el delito. Sin embargo, estos sistemas de IA, si no se diseñan teniendo en cuenta las preocupaciones éticas, pueden terminar exacerbando el sesgo que buscan mitigar, ya sea incorporando directamente factores sesgados o utilizando proxies para el sesgo en sus recomendaciones.[272] Esto puede tener graves consecuencias, incluida la perpetuación de la discriminación contra ciertos grupos.

Por lo tanto, cuando los sistemas de IA son parciales y opacos, plantean preocupaciones con respecto a los estándares de un juicio justo, como la presunción de inocencia, el derecho a ser informado sin demora del origen y la naturaleza de una acusación, el derecho a un juicio justo y la capacidad de defenderse en persona. La opacidad de la toma de decisiones por parte de los sistemas de IA también plantea preocupaciones con respecto a la privación arbitraria de la libertad y el derecho a no ser castigado sin apego a la ley.[273]

> *«El uso de herramientas de evaluación de riesgos para tomar decisiones justas sobre la libertad humana requeriría resolver profundos desafíos éticos, técnicos y estadísticos, incluida la garantía de que las herramientas estén diseñadas y construidas para mitigar el sesgo tanto en el modelo como en las capas de datos, y que existan protocolos adecuados para promover la transparencia y la responsabilidad. Las herramientas actualmente disponibles y bajo consideración para un uso generalizado sufren de varias de estas fallas».*[274]

272 Por ejemplo, según los registros públicos, la policía de Nueva Orleans utilizó un software creado por Palantir para investigaciones criminales de una manera que se extendía más allá del alcance original previsto del software. Tras una secuencia de informes de investigación y una reacción pública significativa, la ciudad rescindió su contrato de seis años con Palantir en marzo de 2018.

273 CAHAI (2020). The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law, disponible en: https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da.

274 Partnership on AI, Report on Algorithmic Risk Assessment Tools in the US Criminal Justice System, disponible en: https://partnershiponai.org/wp-content/uploads/2021/08/Report-on-Algorithmic-Risk-Assessment-Tools.pdfs

Actividad: Los participantes en la capacitación leen una jurisprudencia selecta que trata sobre cajas negras algorítmicas en ADM y sistemas de IA y debaten cómo la IA y los avances tecnológicos afectan los derechos humanos para acceder a los tribunales, a un juicio justo y al debido proceso.[275]

Estado vs. Loomis en los Estados Unidos

En Estado vs. Loomis, la Corte Suprema de Wisconsin determinó que el uso del algoritmo COMPAS, una herramienta patentada de evaluación de riesgos, en la sentencia no violó los derechos de debido proceso del acusado. COMPAS se desarrolló inicialmente para ayudar a las juntas de libertad condicional a determinar el riesgo de reincidencia. Sin embargo, el resultado de COMPAS, una puntuación de evaluación de riesgos, fue utilizado tanto por el Estado como por el tribunal de primera instancia durante la sentencia. Northpointe, Inc., la empresa que creó COMPAS, se negó a revelar su metodología al tribunal o al preso. El tribunal de sentencia le dio al acusado una sentencia de seis años en lugar de libertad condicional, ya que el algoritmo determinó que tenía una probabilidad significativa de reincidencia.[276]

Aunque el Tribunal confirmó la validez de COMPAS, hubo muchas limitaciones impuestas a su solicitud. El algoritmo no se podía utilizar para evaluar si un delincuente cumpliría condena en prisión o para estimar la duración de su condena. Cualquier informe de investigación previo a la sentencia, incluida la puntuación, tenía que incluir un descargo de responsabilidad elaborado de cinco partes sobre las limitaciones del algoritmo. Su uso también requería una justificación separada para la sentencia. La Corte Suprema se negó a aceptar el caso en apelación del acusado.[277]

Sigue siendo una pregunta abierta si es apropiado que el tribunal permita que un algoritmo, en el que los operadores judiciales tienen una visibilidad limitada, desempeñe incluso un papel menor en la privación de libertad de una persona. El fallo de la Corte Suprema de Wisconsin y los documentos de apelación revelan errores fundamentales con respecto al posible funcionamiento de un algoritmo como COMPAS y las protecciones necesarias para que sea útil en la sentencia. Estos malentendidos ofrecen una visión de un marco más prometedor, que permitiría a los algoritmos fortalecer el sistema de justicia sin plantear problemas legales, tecnológicos o éticos.[278]

People vs. Alvin Davis en los Estados Unidos

En este caso, dos testigos afirmaron haber visto a un hombre negro de unos cincuenta años en la propiedad el día anterior al asesinato de una mujer mayor que había sido atacada sexualmente y asesinada allí. En los pocos meses previos al asesinato, docenas de personas, incluido el Sr. Davis y otra persona, habían visitado la residencia de la víctima. El Sr. Davis es un hombre afroamericano que tenía la enfermedad de Parkinson y tenía más de 70 años en el momento del asesinato. Una segunda persona que encajaba en la descripción de los testigos tenía antecedentes de delitos sexuales.

Numerosos sitios y objetos en la escena del crimen fueron muestreados en busca de ADN. Muchos de esos artículos, incluido un bastón que supuestamente se usó para agredir sexualmente a la víctima, no contenían el ADN del Sr. Davis. Aunque STRMix, un software utilizado para el análisis de ADN, pudo hacer coincidir con éxito al Sr. Davis con la muestra de ADN de un cordón de zapato que probablemente se usó para atar a la víctima, el software de ADN tradicional no pudo hacerlo. La fiscalía ponderó ampliamente STRMix ante el jurado. Debido a la enfermedad de Parkinson, el Sr. Davis está confinado a una silla de ruedas. El primer juicio en su contra terminó en un jurado en desacuerdo. Después de un segundo juicio, fue declarado culpable y condenado a cadena perpetua sin libertad condicional.

En People vs. Alvin Davis en California, la Electronic Frontier Foundation (EFF) intervino a favor de la capacidad del Sr. Davis para ver el código fuente de STRMix, el programa de ADN forense que se empleó durante su juicio. La EFF ha afirmado que un acusado tiene derecho a revisar el software de análisis de ADN en varios casos, el más reciente de los cuales es este. En dos de esos casos, Estados Unidos contra Ellis y Estado vs.

275 Grimm P., Grossman M. R., Cormack G. V, Artificial Intelligence as Evidence, Artificial Intelligence as Evidence, 19 Nw. J. Tech. & Intell. Prop. 9, disponible en: https://scholarlycommons.law.northwestern.edu/njtip/vol19/iss1/2

276 Israni E. (2017). Algorithmic Due Process: Mistaken Accountability and Attribution in State v. Loomis, disponible en: https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1.

277 Ibid.

278 Ibid.

Pickett, los tribunales acordaron con EFF que los acusados tenían derecho al código fuente de TrueAllele, uno de los principales rivales de STRMix.[279]

Para garantizar que el resultado del software de coincidencia de ADN utilizado contra ellos sea preciso, se debe permitir que los acusados revisen cómo funciona el software. Dado que puede haber fallas de codificación, tener acceso al código fuente no puede ser un sustituto del testimonio sobre cómo debe funcionar el software. Esto es particularmente cierto para el software de ADN forense más reciente, como STRMix y TrueAllele, que está plagado de problemas de precisión y confiabilidad.[280] En realidad, STRMix se examinó previamente y se encontró que contenía fallas de programación que pueden haber derivado en resultados erróneos en 60 casos en Queensland, Australia.[281]

Estado de Nueva Jersey vs. Pickett; Estados Unidos vs. Ellis

Tanto en Nueva Jersey vs. Pickett[282] como en Estados Unidos vs. Ellis[283] , la defensa buscó acceso al código fuente del software de una empresa (TrueAllele). TrueAllele se utiliza para realizar un estudio de genotipado probabilístico en muestras de ADN. Los tribunales en ambos casos concluyeron que el acceso al código debería concederse a la defensa supeditada a una orden de protección. El tribunal de Pickett enfatizó que «cualquier cosa que no sea el acceso total contraviene los principios fundamentales de equidad, lo que indudablemente compromete el derecho del acusado a presentar una defensa completa». Si bien estas herramientas son diferentes de las tecnologías de IA basadas en datos, las resoluciones que muestran que se puede acceder al código fuente del software en los procedimientos penales establecen un precedente alentador para otras tecnologías avanzadas que reclaman protecciones de secretos comerciales.[284]

Estado de Nueva Jersey vs. Francisco Arteaga en los Estados Unidos

New Jersey vs. Arteaga es un ejemplo de un caso que destaca la importancia de la detectabilidad de los algoritmos de IA y sus entradas de datos en los casos judiciales. En 2019, un negocio en West New York, Nueva Jersey, fue asaltado usando una pistola, y Francisco Arteaga fue posteriormente identificado como el sospechoso y acusado del robo. Antes de la identificación del Sr. Arteaga, la policía de Nueva Jersey descubrió que los testigos en el lugar del crimen no pudieron identificar al delincuente, y una búsqueda de reconocimiento facial realizada por el Centro Regional de Inteligencia de Operaciones de Nueva Jersey no arrojó resultados.

Después de este intento fallido de identificación de los sospechosos, el Departamento de Policía de Nueva York realizó una búsqueda de reconocimiento facial utilizando fotos fijas cortadas de cámaras de vigilancia a nivel de calle. El Sr. Arteaga estaba entre los resultados de búsqueda, y el analista de reconocimiento facial de la policía de Nueva York lo identificó como la «posible coincidencia». Posteriormente, la policía colocó la foto del Sr. Arteaga en una alineación de fotos, donde dos testigos finalmente lo identificaron, a pesar de los procesos defectuosos utilizados para realizar las alineaciones. A pesar de la importancia del emparejamiento basado en algoritmos para el caso, la defensa no recibió ninguna información sobre el algoritmo que lo generó. El Sr. Arteaga exigió el descubrimiento de la tecnología de reconocimiento facial utilizada por la policía de Nueva York, la foto original y cualquier edición realizada por la policía de Nueva York antes de realizar una búsqueda, e información sobre el analista que realizó la búsqueda que lo identificó. El tribunal de distrito de Nueva Jersey rechazó su petición de ordenar el descubrimiento.

EPIC junto con la Electronic Frontier Foundation (EFF) y la National Association of Criminal Defense Lawyers (NACDL) presentaron un escrito informando al tribunal sobre cómo se producen los errores en los sistemas de reconocimiento facial, el potencial de sesgo en esos sistemas. Argumentaron que el descubrimiento es la última oportunidad para corregir estos errores. El informe describía la secuencia de procedimientos necesarios para realizar una búsqueda de reconocimiento facial, todos los cuales implican decisiones humanas que pueden agregar inexactitudes y aumentar la probabilidad de identificación errónea. El escrito sostiene que la revisión humana después de una búsqueda no puede considerarse un remedio para los errores algorítmicos.[285] El caso está ahora ante el Juez del Tribunal de Apelación.[286]

279 Zhao H. (2021). EFF tells California Court that Forensic

280 Zhao H. (2021). How Your DNA—or Someone Else's-Can Send You to Jail, disponible en: https://www.eff.org/deeplinks/2021/05/how-your-dna-or-someone-elses-can-send-you-jail.

281 Murray D. (2015). Queensland authorities confirm 'miscode' affects DNA evidence in criminal cases, disponible en: http://www.couriermail.com.au/news/queensland/queensland-authorities-confirm-miscode-affects-dna-evidence-in-criminal-cases/news-story/833c580d3f1c59039efd1a2ef55af92bc

282 Estado de Nueva Jersey vs. Corey Pickett, disponible en: https://law.justia.com/cases/new-jersey/appellate-division-published/2021/a4207-19.html.

283 EFF, Estados Unidos vs. Ellis, disponible en: https://www.eff.org/cases/united-states-v-ellis

284 Grupo de trabajo de la NACDL sobre vigilancia predictiva (2021). Garbage in, gospel out. How Data-Driven Policing Technologies Entrench Historic Racism and 'Tech-wash' Bias in the Criminal Legal System, disponible en: https://www.nacdl.org/Document/GarbageInGospelOutDataDrivenPolicingTechnologies

285 EPIC Amicus Brief, New Jersey vs. Arteaga, disponible en: https://epic.org/documents/new-jersey-v-arteaga/

286 Murphy R. (2022). Lawyers and digital rights advocates want the facial recognition process exposed in court, disponible en: https://localtoday.news/nj/lawyers-and-digital-rights-advocates-want-the-facial-recognition-process-exposed-in-court-52064.html

Una de las mayores amenazas generadas por el uso de sistemas de IA en la administración de justicia es el llamado sesgo de automatización, que es la tendencia de los humanos a recibir sin crítica la solución que ofrece la inteligencia artificial como correcta, produciendo una validación automática por parte de los humanos. Este es un riesgo particularmente aberrante en la administración de justicia, ya que puede conducir a una confianza ciega en las decisiones propuestas por el sistema, a considerar que la única jurisprudencia existente es la propuesta por la máquina o a considerar correcta una evaluación de la posibilidad de reincidencia. Con el tiempo, esto llevaría a un cambio en el razonamiento de las decisiones dirigidas a justificar por qué no se sigue el resultado que ofrece el sistema, posibilidad que se ve agravada por la desproporcionada carga de trabajo de la mayoría de nuestros tribunales, lo que lleva a un sistema de trabajo en el que la cantidad y la rapidez priman sobre la calidad. Es por esta razón que la desviación del juez de cualquier decisión, ya sea asistida o automatizada, no puede dar lugar a ningún tipo de represalia, sanción, inspección o régimen disciplinario. Si prevalece la supervisión y el control humanos, el control debe ser efectivo. Preguntas clave que debemos hacernos al respecto:

(i) ¿Cómo afecta la resolución de un caso por parte de un sistema de IA, en lugar de un juez, al derecho efectivo de acceso a los tribunales, a un juicio justo y al debido proceso?

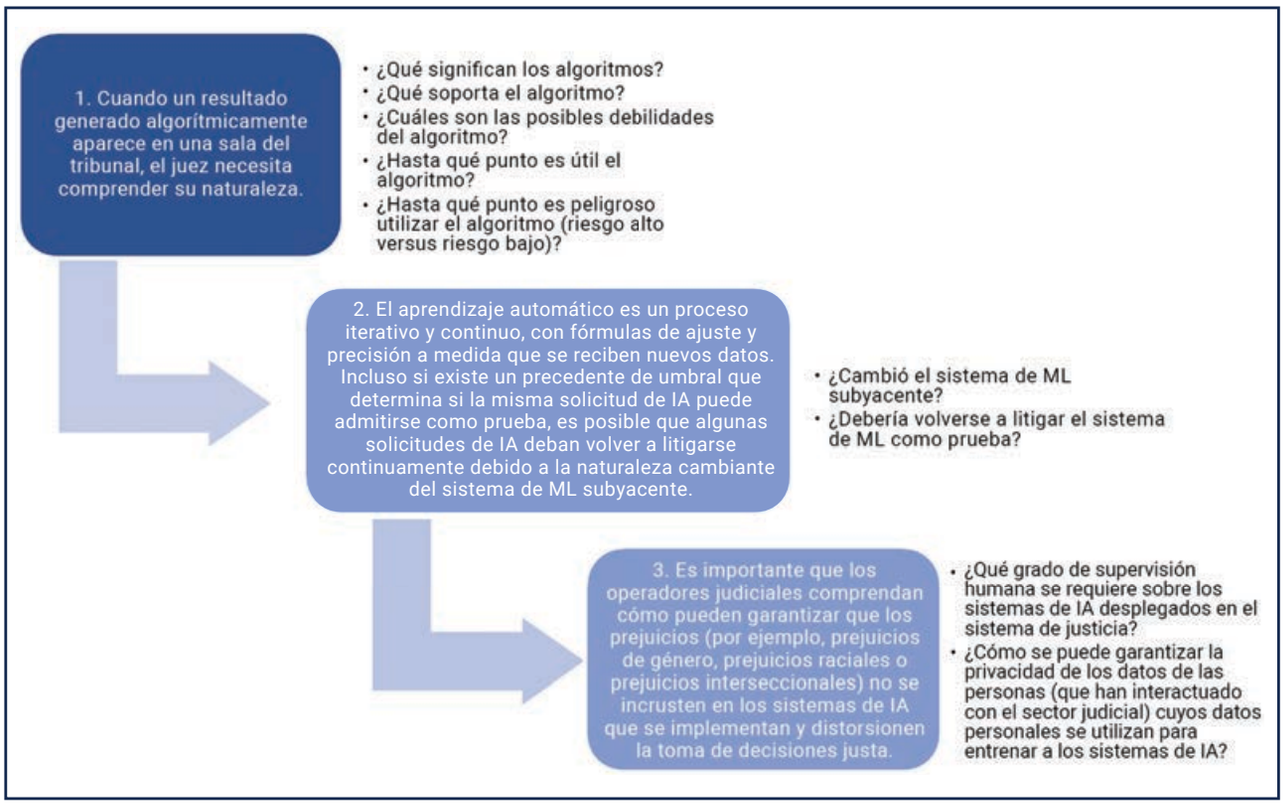
(ii) ¿Cómo se articulará la motivación de las decisiones judiciales? Los ciudadanos tienen derecho a conocer la motivación de las sentencias y los jueces tienen el deber de motivarlas. En el caso de una caja negra, el razonamiento lógico de la conclusión no es transparente ni se puede obtener.

(iii) En el caso de la existencia de una propuesta de proyecto de decisión o la aplicación de la jurisprudencia por un sistema de IA que alimenta una decisión/sentencia tomada en última instancia por un juez humano, ¿tienen las partes derecho a conocer el razonamiento del sistema de IA, y podría utilizarse ese argumento como motivo de apelación o como argumento para apoyar la apelación? El derecho a la transparencia del algoritmo y las deliberaciones secretas del tribunal son dos cuestiones separadas que no deben confundirse.

Los sistemas de IA deben ser tenidos en cuenta como herramientas auxiliares y de apoyo, sin atribuirles un valor decisivo o caer en la sobreestimación, sin olvidar la necesaria motivación judicial y la esencial individualización de las penas. Debe garantizarse el derecho a no ser objeto de una decisión únicamente automatizada, el derecho a ser informado de la decisión automatizada, el derecho a impugnar o revisar las decisiones automatizadas o algorítmicas y el derecho a solicitar supervisión e intervención humanas.

La Figura 14 a continuación describe algunos pasos que los operadores judiciales podrían seguir al decidir casos que involucran a la IA y los derechos humanos:

Figura 14. Pasos que los operadores judiciales podrían seguir al decidir casos que involucran a la IA y los derechos humanos



Fuente: Autores

Actividad: Garantizar que los sistemas de IA se utilicen de una manera que respete los principios de un juicio justo es fundamental para mantener la integridad del sistema legal. Aquí hay un ejemplo de caso hipotético que ilustra la importancia de la IA para garantizar un juicio justo. Revisen los hechos del caso y debatan qué leyes se habrían aplicado si el caso se hubiera juzgado en su jurisdicción. ¿Cuál habría sido el resultado del caso?

Título del caso: El Estado vs. Pedro Pérez

Antecedentes: Pedro Pérez enfrenta cargos penales relacionados con un robo que ocurrió en una tienda. La fiscalía se basa en las imágenes de las cámaras de vigilancia como una pieza clave de evidencia. La defensa, sin embargo, argumenta que las imágenes no son concluyentes y que Pedro Pérez está siendo acusado injustamente.

Papel de la IA para garantizar un juicio justo:

1. IA de análisis de vídeo: la fiscalía introduce un sistema de análisis de vídeo basado en IA que pretende mejorar y analizar las imágenes de vigilancia. Se dice que este sistema de IA tiene la capacidad de identificar rostros, mejorar la calidad de la imagen y detectar comportamientos sospechosos.
2. Preocupaciones expresadas por la defensa: la defensa plantea preocupaciones sobre la precisión y los posibles sesgos del sistema de IA. Argumentan que la IA puede haber sido entrenada en conjuntos de datos sesgados y que sus resultados pueden no ser confiables.
3. Testigos expertos: tanto la fiscalía como la defensa llaman a testigos expertos para que testifiquen sobre las capacidades y limitaciones del sistema de IA. El testigo experto de la defensa cuestiona la precisión de la IA y destaca los posibles sesgos.
4. Transparencia y explicabilidad: la defensa solicita que los algoritmos y procesos de toma de decisiones del sistema de IA sean divulgados para su examen. Argumentan que sin transparencia y explicabilidad, no se puede confiar en los hallazgos de la IA.
5. Revisión independiente: el tribunal ordena una revisión independiente de los resultados y algoritmos del sistema de IA por parte de un tercero neutral. Esta revisión tiene como objetivo evaluar la precisión y la imparcialidad de los hallazgos de la IA.
6. Precedente legal: el caso llama la atención sobre la necesidad de estándares y pautas legales con respecto al uso de la IA en los juicios penales. El tribunal considera si el uso de la IA en este caso cumple con los estándares legales existentes y los principios de equidad.

Resultado:

En última instancia, el tribunal decide admitir el análisis de video mejorado con IA como prueba, pero con condiciones:

- Los algoritmos y los procesos de toma de decisiones del sistema de IA deben divulgarse a la defensa y al revisor independiente.
- El tribunal reconoce que los sistemas de IA pueden introducir sesgos y errores y que se permitirá el testimonio de expertos con respecto a las limitaciones de la IA.
- El revisor independiente evaluará los hallazgos de la IA y proporcionará un informe al tribunal.

Este caso hipotético destaca la importancia de la transparencia, la equidad y la responsabilidad al utilizar la IA en el sistema legal. También subraya la necesidad de normas y directrices legales para garantizar que las tecnologías de IA no comprometan los principios de un juicio justo, incluido el derecho a la defensa, el derecho a impugnar las pruebas y el derecho a interrogar y contrainterrogar a los testigos.

Recurso efectivo

El uso de sistemas de IA en situaciones en las que los derechos humanos están en juego puede presentar dificultades para garantizar el derecho a la reparación. Dado que muchos sistemas de IA son opacos, las personas pueden desconocer cómo se tomaron las decisiones que afectan sus derechos o si el proceso fue discriminatorio. A menudo, el operador judicial que utiliza el sistema de IA puede ser incapaz de explicar el proceso automatizado de toma de decisiones. Estos problemas se ven agravados por el despliegue de sistemas de IA que recomiendan, toman o hacen cumplir las decisiones dentro del poder judicial, las mismas instituciones responsables de proteger los derechos, incluido el derecho a un recurso efectivo.[287]

Refutabilidad

Las personas y los grupos afectados deben disponer de medios eficaces para impugnar las determinaciones y decisiones pertinentes. Como condición previa necesaria, la existencia, el proceso, la justificación, el razonamiento y el posible resultado de los sistemas algorítmicos a nivel individual y colectivo deben explicarse y aclararse de manera oportuna, imparcial, fácil de leer y accesible para las personas cuyos derechos o intereses legítimos puedan verse afectados, así como para las autoridades públicas pertinentes. La impugnación debe incluir la oportunidad de ser escuchado, una revisión exhaustiva de la decisión y la posibilidad de obtener una decisión no automatizada. Este derecho no puede ser renunciado, y debe ser asequible y fácilmente exigible antes, durante y después del despliegue, incluso mediante la provisión de puntos de contacto y líneas directas de fácil acceso.

Fuente: Recomendación CM/Rec (2020) del Consejo de Europa del Comité de Ministros a los Estados miembros sobre los impactos de los sistemas algorítmicos en los derechos humanos (Adoptada por el Comité de Ministros el 8 de abril de 2020 en la 1373.ª reunión de los Delegados de los Ministros)

Los procesos automatizados de toma de decisiones se prestan a desafíos para la capacidad de las personas de obtener una solución efectiva. Estos incluyen la opacidad de la decisión en sí, su base y si las personas han dado su consentimiento para el uso de sus datos al tomar esta decisión o incluso son conscientes de cómo les afecta. No está claro a quién deben expresar las personas sus problemas con la decisión debido a la dificultad para asignar la responsabilidad de la decisión. Debido a la naturaleza de los juicios que se realizan automáticamente, sin o con poca participación humana, y con un enfoque en la eficiencia en lugar del razonamiento humano-contextual, las organizaciones que implementan sistemas ADM tienen una obligación aún mayor de proporcionar a las personas afectadas un método para buscar reparación.[288] En este contexto, vale la pena mencionar que la propuesta de directiva de responsabilidad de IA de la UE crearía una "presunción de causalidad" refutable, para aliviar la carga de la prueba para establecer el daño causado por un sistema de IA. Esto aliviará algunos de los obstáculos al presentar una reclamación por daños causados por un sistema de IA. Además, daría a los tribunales nacionales la facultad de ordenar la divulgación de pruebas sobre sistemas de IA sospechosos de haber causado daños.[289]

287 Declaración de Toronto, disponible en: https://www.torontodeclaration.org/declaration-text/english/

288 Comité de expertos en intermediarios de Internet (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Council of Europe Study, DGI/2017/12, disponible en: https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

289 Proposal for a Directive on adapting non contractual civil liability rules to artificial intelligence, disponible en: https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en

Actividad: Los participantes en la capacitación leen una jurisprudencia selecta que trata sobre cajas negras algorítmicas en ADM y sistemas de IA y debaten cómo la IA y los avances tecnológicos afectan el derecho a la reparación.

People vs. Chubbs (2015) en los Estados Unidos

Un Tribunal de Apelaciones de California confirmó un privilegio probatorio de secreto comercial en un caso penal en 2015 para evitar la divulgación del código fuente de TrueAllele a la defensa. Se hace referencia a la sentencia en el caso People vs. Chubbs en los EE. UU. para negar a los acusados el acceso a pruebas de secretos comerciales.[290] El tribunal dictaminó que un acusado no tiene derecho al código fuente de un algoritmo de ADN utilizado para identificar al acusado, prima facie. El propietario de un secreto comercial tiene derecho a negarse a divulgar el secreto si la concesión de ese derecho no servirá para ocultar el fraude o promover la injusticia.[291] En este caso, el Tribunal de Apelaciones de California extendió un privilegio probatorio de secreto comercial en un caso penal. Permitió al desarrollador retener "por completo" el código fuente. El caso Chubbs ha formado la base de un nuevo cuerpo de jurisprudencia en los Estados Unidos que niega el acceso al código fuente subyacente de los algoritmos utilizados en todo el sistema de justicia penal.[292]

Caso de Uber sobre el uso del programa de detección de fraudes "Mastermind" en Europa

Un caso reciente contra Uber se ha basado en el artículo 22 del RGPD, que establece que las personas "tienen derecho a no estar sujetas a una decisión basada únicamente en el procesamiento automatizado, incluida la elaboración de perfiles, que produzca efectos legales que les conciernan o que les afecte significativamente de manera similar".[293] Los solicitantes solicitaron que el Tribunal de Distrito de Ámsterdam analice Mastermind, el sofisticado programa de detección de fraudes de Uber.

Al invocar las protecciones del RGPD contra la toma de decisiones automatizada, los conductores de Uber en el Reino Unido y Portugal afirmaron que fueron despedidos injustamente por el algoritmo antifraude de la empresa. Los solicitantes afirmaron que el algoritmo utilizado por Uber fue automatizado (sin intervención humana significativa) y resultó en la terminación de su trabajo con Uber. Sin darles la posibilidad de impugnar la decisión tomada por la empresa.[294]

El propósito declarado de Mastermind es ayudar a Uber a vigilar eficazmente su plataforma. La demanda afirma que Uber no ha demostrado que su personal tenga el conocimiento suficiente sobre los aportes a su sistema de lucha contra el fraude para pronosticar el resultado o explicar los juicios del algoritmo. También indicó que Uber debe proporcionar a los socios de la App información precisa sobre cualquier presunta infracción. Según la queja, las cartas de desactivación de Uber eran en su mayoría genéricas y omitían información sobre el presunto fraude. Además, los conductores no tuvieron la oportunidad de refutar las acusaciones.[295]

290 Milner-Smith H., Copper D. (2017). When a computer program keeps you in jail, disponible en: https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-criminal-justice.html

291 El pueblo vs. Tribunal Superior del Condado de Los Ángeles (Chubbs) (Cal. Ct. App. 2015), disponible en: https://www.quimbee.com/cases/people-v-chubbs

292 Chaney G. (2019). The Criminal Justice System's Algorithms Need Transparency, disponible en: https://www.law360.com/articles/1143086/the-criminal-justice-system-s-algorithms-need-transparency

293 https://ekker.legal/wp-content/uploads/2020/10/Court-request-Uber-account-deactivation-unofficial-translation.pdf. Artículo 22 del RGPD: "El interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos que le conciernan o que le afecten significativamente de manera similar".

294 Huseinzade N. (2021). Algorithm Transparency: How to Eat the Cake and Have It Too, disponible en: https://europeanlawblog.eu/2021/01/27/algorithm-transparency-how-to-eat-the-cake-and-have-it-too/

295 Claburn T. (2020). Uber drivers take ride biz to European court over 'Kafkaesque' algorithmic firings by Mastermind code, disponible en: https://www.theregister.com/2020/10/26/uber_algorithmic_lawsuit/

Un tribunal de distrito de Ámsterdam ordenó a Uber que reintegre a los socios de la App que fueron despedidos injustamente por el algoritmo de la empresa. También ha ordenado a Uber que indemnice a los socios de la App con más de €100 000 en daños.[296]

El caso de Robodebt en Australia

En 2016, el gobierno australiano introdujo "Robodebt", un sistema automatizado de comparación de datos para reemplazar el examen humano de los datos de ingresos de los beneficiarios de asistencia social. El objetivo era detectar sobrepagos o fraudes. Sin embargo, las personas marcadas por el algoritmo como sospechosas debían proporcionar pruebas para demostrar su inocencia a través de un formulario en línea o arriesgarse a perder sus beneficios por completo. Este proceso ha tenido efectos perjudiciales en muchas personas.

El algoritmo, sin embargo, tomó datos de la autoridad fiscal (que se basan en un año completo) y los comparó con los ingresos quincenales, ignorando el hecho de que los ingresos de los beneficiarios de asistencia social a menudo son muy irregulares, debido a contratos a corto plazo o trabajo temporal. Como resultado, miles de personas fueron privadas erróneamente de los pagos de asistencia social, y muchas de ellas no pudieron impugnar estas decisiones ya que las notificaciones automatizadas se enviaron a una dirección antigua o no tuvieron acceso al portal a través del cual podrían haber enviado las pruebas requeridas.

En muchos casos, las personas de repente se enfrentaron en serias deudas, e incluso se informaron algunos casos de suicidio. Algunas fuentes calculan que las autoridades han intentado reclamar casi 600 millones de AUD (360 millones de euros) a los ciudadanos basándose en este sistema, que a menudo generaba errores, pero bajo el cual la carga de la prueba pasaba al individuo. Los resultados fueron muy difíciles de cuestionar. Este caso ha reavivado el debate sobre cómo se utilizan los algoritmos y la coincidencia de datos para fundamentar las decisiones.[297]

El acuerdo propuesto para una demanda colectiva contra el Commonwealth de Australia con respecto a su uso de Robodebt fue aprobado por el Tribunal Federal el 11 de junio de 2021. Según el acuerdo, el Estado Libre Asociado pagará $112 millones (incluidos los costos legales) a ciertos miembros del grupo como intereses, se abstendrá de recaudar, exigir o recuperar cualquier deuda inválida de ciertos miembros del grupo y aceptará declaraciones judiciales de que algunas de sus decisiones administrativas no se tomaron válidamente.[298]

296 Nawrat A. (2021). HR tech gone wrong? Uber told to reinstate drivers after 'robo-firing', disponible en: https://www.unleash.ai/hr-technology/court-rules-against-uber-robo-firing-employee-surveillance/.
297 Human Rights Law Centre (2021). The Federal Court approves a $112 million settlement for the failures of the Robodebt system, disponible en: https://www.hrlc.org.au/human-rights-case-summaries/2021/9/30/the-federal-court-approves-a-112-million-settlement-for-the-failures-of-the-robodebt-system.
298 Ibid. Véase también: Katherine Prygodicz & Ors v The Commonwealth of Australia (No 2) [2021] FCA 634 (11 de junio de 2021).

"Todas las personas son iguales ante la ley y tienen derecho sin discriminación a igual protección de la ley. A este respecto, la ley prohibirá toda discriminación y garantizará a todas las personas protección igual y efectiva contra cualquier discriminación por motivos de raza, color, sexo, idioma, religión, opiniones políticas o de cualquier índole, origen nacional o social, posición económica, nacimiento o cualquier otra condición social".

– Artículo 26 del PIDCP

"En los Estados en que existan minorías étnicas, religiosas o lingüísticas, no se negará a las personas que pertenezcan a dichas minorías el derecho que les corresponde, en común con los demás miembros de su grupo, a tener su propia vida cultural, a profesar y practicar su propia religión y a emplear su propio idioma".

– Artículo 27 del PIDCP

"Los Estados partes en el presente Pacto se comprometen a garantizar a hombres y mujeres la igualdad en el goce de todos los [...] derechos enunciados en el presente Pacto".

– Artículo 3 del PIDCP y del PIDESC

Los derechos a la protección contra la discriminación pueden ser violados por los sistemas de IA, debido a (i) el potencial de sesgo por parte de los desarrolladores de algoritmos; (ii) sesgo incrustado en el modelo sobre el cual se construyen los sistemas de IA; (iii) sesgo incrustado en los conjuntos de datos utilizados para entrenar los modelos; o (iv) sesgo introducido cuando dichos sistemas se aplican en entornos del mundo real. Estos riesgos se agravan en situaciones en las que se implementan sistemas de IA para ayudar a los operadores judiciales en sus actividades cotidianas.

El diseño de los sistemas de IA y su uso en los procedimientos judiciales deben regirse con el objetivo de producir resultados respetuosos de los derechos humanos y no discriminatorios. Se deben establecer normas y salvaguardias mínimas; si no se pueden cumplir, no se debe utilizar el sistema de IA en cuestión.

Además, la IA debe regularse para que sea lo suficientemente transparente y explicable como para permitir una revisión independiente efectiva. El diseño y despliegue de sistemas de IA debe cumplir y hacer efectivo el derecho de acceso a los tribunales, el derecho a la presunción de inocencia y el derecho a la libertad, entre otros.

Ningún ser humano debe estar expuesto a una decisión automatizada que resulte en antecedentes penales, y las tecnologías de IA no deben comprometer el derecho a un juicio justo por parte de un tribunal imparcial e independiente. Los sistemas de IA no deben etiquetar previamente a las

299 Fair Trials, Regulating Artificial Intelligence for Use in Criminal Justice Systems in the EU Policy Paper, disponible en: https://www.fairtrials.org/sites/default/files/Regulating%20Artificial%20Intelligence%20for%20Use%20in%20Criminal%20Justice%20Systems%20-%20Fair%20Trials.pdf

personas como delincuentes sin juicio, ni deben permitir a las autoridades tomar medidas injustificadas y desproporcionadas contra las personas sin una sospecha razonable.

Cuando los sistemas de IA informan las decisiones sobre privaciones de libertad, deben ajustarse para crear resultados que favorezcan la liberación, y no deben facilitar la detención, excepto como último recurso. Para garantizar que los sistemas de IA logren el efecto deseado de reducir las tasas de detención preventiva, deben someterse a pruebas rigurosas.[299]

Cuestiones que deben ser tenidas en cuenta por los operadores judiciales al evaluar el impacto potencial y el riesgo de la IA en los derechos a la protección contra la discriminación

- ¿Cómo, en todo caso, podría el sistema de IA resultar en discriminación, tener impactos discriminatorios en los titulares de derechos o funcionar de manera diferenciada para diferentes grupos de manera discriminatoria o perjudicial?
- ¿Cómo podría el uso del sistema de IA exacerbar las desigualdades o la discriminación existentes en las poblaciones a las que afecta?
- ¿De qué otras maneras, si las hubiera, podría el uso de este sistema contribuir o exacerbar la inequidad o la desigualdad?

Fuente: Leslie D., Burr C., Aitken M., Cowls J., Katell M., Briggs M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer, The Council of Europe, disponible en: https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf.

Los sistemas de IA deben diseñarse para garantizar que no produzcan resultados discriminatorios, asegurando que las personas sospechosas y acusadas no estén en desventaja, ya sea directa o indirectamente, en función de sus características, como la raza, el origen étnico, la nacionalidad, la minoría o el nivel socioeconómico. Los sistemas de IA deben estar sujetos a pruebas obligatorias antes y después del despliegue para identificar y corregir cualquier efecto discriminatorio. Consulte el Módulo 3 que analiza el sesgo algorítmico en detalle.[300]

Los sistemas de IA deben ser transparentes y comprensibles para que sus usuarios clave, como los responsables de la toma de decisiones, las partes en un litigio y los acusados, puedan comprenderlos y analizarlos. Los intereses comerciales o de propiedad, como los secretos comerciales, deben equilibrarse con los requisitos relacionados con la transparencia. Cada sistema de IA debe poder auditarse por un auditor independiente, y sus procesos deben ser replicables para este propósito.[301]

300 Ibid.
301 Ibid.

Actividad: Los participantes en la capacitación leen los hechos de los casos de Deliveroo y Foodhino y analizan cómo la opacidad de los algoritmos de IA y su funcionamiento como cajas negras afectan los derechos a la protección contra la discriminación y el sesgo.

Caso Deliveroo (2021)[302]

Deliveroo es un servicio de entrega de alimentos que funciona como un mercado en tres dimensiones a través de una aplicación en línea. Conecta a consumidores locales, restaurantes, tiendas de comestibles y personas usuarias. Tres sindicatos entraron en controversia con Deliveroo en los tribunales italianos por violar las leyes laborales regionales. En este caso, el tribunal de Bolonia dictaminó que el algoritmo de calificación de reputación de Deliveroo discriminaba a los repartidores de comida o a los conductores.[303] Al parecer, el algoritmo ML examinado por el tribunal se utilizó para estimar la "fiabilidad" de un conductor. El tribunal señaló que el "índice de fiabilidad" del conductor se vería afectado si no cancelaba un turno reservado con antelación utilizando la app al menos 24 horas antes de la hora de inicio. Dado que el algoritmo priorizó ofrecer turnos en bloques de tiempo de alta demanda a usuarios más confiables, las personas usuarias que no puedan cumplir con sus turnos, incluso en caso de una emergencia o enfermedad grave, tendrán menos opciones de trabajo en el futuro. Según el tribunal, el hecho de que el algoritmo ML no considerara la causa de una cancelación constituía discriminación y penalizaba injustamente a los conductores que tenían razones legalmente válidas para no trabajar. Se ordenó a Deliveroo que indemnizara a los demandantes con €50.000.[304]

El tribunal también señaló que los criterios para el funcionamiento del algoritmo no estaban definidos en la aplicación más allá de los aspectos genéricos de confiabilidad y participación, ni fueron suministrados al tribunal por la corporación demandada, lo que obstaculizó una evaluación exhaustiva del asunto.[305]

Caso Foodinho (2021)

Foodinho, otro servicio de entrega de alimentos con sede en Italia, fue penalizado con 2,6 millones de euros por la Autoridad Italiana de Protección de Datos (Garante) por utilizar algoritmos de medición de rendimiento discriminatorios en relación con sus empleados. La autoridad declaró a Foodinho como transgresora de los principios de transparencia, seguridad y privacidad por defecto y por diseño, y responsabilizó a la empresa por no tomar las medidas adecuadas para proteger los derechos y libertades de sus empleados (es decir, los conductores) del ADM discriminatorio. En términos de gestión algorítmica de los trabajadores por encargo, la decisión del Garante es la primera de su tipo. El Garante alegó que la dirección de Foodinho había violado el artículo 22(3) del RGPD.[306]

302 Colossa A. (2021). Algorithms, biases, and discrimination in their use: About recent judicial rulings on the subject, disponible en: https://www.ciat.org/ciatblog-algorithms-biases-and-discrimination-in-their-use-about-recent-judicial-rulings-on-the-subject/?lang=en

303 Lomas N. (2021). Italian court rules against 'discriminatory' Deliveroo rider-ranking algorithm, disponible en: https://techcrunch.com/2021/01/04/italian-court-rules-against-discriminatory-deliveroo-rider-ranking-algorithm/.

304 Geiger G. (2021). Court Rules Deliveroo Used 'Discriminatory' Algorithm, disponible en: https://www.business-humanrights.org/en/latest-news/court-rules-deliveroo-used-discriminatory-algorithm/.

305 Ibid.

306 Milner-Smith et al. (2020). Italian Supervisory Authority Fines Foodinho Over Its Use of Performance Management Algorithms, disponible en: https://www.insideprivacy.com/gdpr/italian-supervisory-authority-fines-foodinho-over-its-use-of-performance-management-algorithms/.

En su fallo, el Garante ha declarado que Foodinho participa en dos tipos diferentes de actividades de procesamiento automatizado: una se encuadra dentro del ámbito del "sistema de excelencia" y la otra es un componente del sistema que distribuye pedidos en función de un algoritmo interno conocido como "Jarvis". El método de puntuación interna utilizado por Foodinho para proporcionar franjas horarias de entrega a sus conductores se conoce como el "sistema de excelencia", que califica a cada conductor. Los socios de la App con calificaciones más altas tienen prioridad a la hora de determinar las franjas horarias de entrega. En la práctica, esto significa que los conductores "menos excelentes" están excluidos de la asignación de franjas horarias de entrega si los conductores "más excelentes" ya han tomado todas las franjas horarias de entrega disponibles. El "puntaje de excelencia" está determinado por un proceso estadístico automatizado que tiene en cuenta principalmente los comentarios de los clientes y socios comerciales, así como las tasas de entrega. Es importante destacar que los comentarios positivos tienen menos peso que los negativos y el sistema penaliza a los conductores que no alcanzan los niveles de entrega requeridos. El algoritmo ("Jarvis") que asigna pedidos utiliza información que incluye el paradero geográfico de las personas usuarias según lo determinado por sus dispositivos GPS, la ubicación de recogida, la dirección de entrega, los requisitos de pedidos especiales y el tipo de vehículo utilizado. Jarvis asigna pedidos y automatiza completamente el procesamiento de estos datos. Sin embargo, Foodinho no explicó específicamente al Garante cómo este algoritmo está vinculado al sistema de excelencia.[307]

Libertad de expresión y acceso a la información

"Toda persona tiene derecho a la libertad de pensamiento, de conciencia y de religión. Este derecho incluye la libertad de tener o de adoptar la religión o las creencias de su elección, así como la libertad de manifestar su religión y sus creencias, individual o colectivamente, tanto en público como en privado, mediante el culto, la celebración de los ritos, las prácticas y la enseñanza. Nadie estará sujeto a coerción que pueda perjudicar su libertad de tener o adoptar una religión o creencia de su elección".

– Artículo 18 del PIDCP y artículo 18 de la DUDH

"Todo individuo tendrá el derecho de sostener opiniones sin interferencia. Toda persona tiene derecho a la libertad de expresión; este derecho comprende la libertad de buscar, recibir y difundir informaciones e ideas de toda índole, sin consideración de fronteras, ya sea oralmente, por escrito o en forma impresa o artística o por cualquier otro procedimiento de su elección."

– Artículo 19 del PIDCP

Varios marcos jurídicos internacionales y principios rectores establecen que los derechos humanos a la libertad de expresión y el acceso a la información se extienden a Internet. En 2011, el Comité de Derechos Humanos de la ONU emitió el Comentario General N.º 34[308] afirmando que el Artículo 19

307 Ibid.
308 ONU (2011). Comentario general N.º 34, disponible en: https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf
309 ONU (1976). Pacto Internacional de Derechos Civiles y Políticos, disponible en: https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights

del Pacto Internacional de Derechos Civiles y Políticos[309] protege todas las formas de expresión y los medios de su difusión, incluidas todas las formas de expresión electrónicas y basadas en Internet (incluido el acceso a la información en línea). Esto significa que el principio de salvaguardar el derecho a la libertad de expresión se extiende al espacio en línea tal como lo hace en el mundo fuera de línea.[310] En 2012, el Consejo de Derechos Humanos de la ONU adoptó una innovadora Resolución 20/8[311] para promover, proteger y garantizar el disfrute de los derechos humanos en línea. Esta resolución afirma la importancia de defender los derechos humanos en la era digital: "Los mismos derechos que las personas tienen fuera de línea también deben protegerse en línea, en particular la libertad de expresión, que es aplicable independientemente de las fronteras y a través de cualquier medio de su elección, de conformidad con los artículos 19 de la Declaración Universal de Derechos Humanos y el Pacto Internacional de Derechos Civiles y Políticos.[312] Del mismo modo, la resolución del Consejo de Derechos Humanos de la ONU de 2018 sobre la promoción, protección y disfrute de los derechos humanos en Internet declaró que "los mismos derechos que las personas tienen fuera de línea también deben protegerse en línea, en particular la libertad de expresión"[313] e instó a todos los Estados a garantizar estos derechos.

Los informes anuales y temáticos del relator especial abordan diversos temas, como la vigilancia estatal de las comunicaciones[314], la salvaguardia de los derechos de los ciudadanos durante las elecciones[315], el discurso de odio en línea[316] el cifrado y el anonimato[317], el derecho de los niños a expresarse[318], el papel del sector privado[319] y los proveedores de acceso digital[320], el impacto de la inteligencia artificial en los derechos de los ciudadanos[321], la protección de la libertad de expresión de los periodistas[322] y la prevención de la censura al abordar el abuso de género en línea[323].

310 ONU (2011). Comentario general N.º 34 (párr. 15), disponible en: https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf
311 Consejo de Derechos Humanos de las Naciones Unidas (2021). The promotion, protection and enjoyment of human rights on the Internet, disponible en: https://ap.ohchr.org/documents/dpage_e.aspx?si=a/hrc/res/20/8
312 ONU (AGNU) (2012). The promotion, protection and enjoyment of human rights on the Internet, 16 de julio de 2012, A/HRC/RES/20/8, disponible en: http://ap.ohchr.org/documents/dpage_e.aspx?si=a/hrc/res/20/8
313 Consejo de Derechos Humanos de las Naciones Unidas (2018). The Promotion, Protection and Enjoyment of Human Rights on the Internet, disponible en: https://digitallibrary.un.org/record/1639840
314 Consejo de Derechos Humanos de las Naciones Unidas (2013). Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión, disponible en: https://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A.HRC.23.40_EN.pdf
315 Consejo de Derechos Humanos de las Naciones Unidas (2014). Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G14/071/50/PDF/G1407150.pdf?OpenElement
316 Asamblea General de la ONU (2019). Promotion and protection of the right to freedom of opinion and expression, disponible en: https://www.ohchr.org/Documents/Issues/Opinion/A_74_486.pdf
317 Consejo de Derechos Humanos de las Naciones Unidas (2015). Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G15/095/85/PDF/G1509585.pdf?OpenElement
318 Asamblea General de la ONU (2014). Promoción y protección del derecho a la libertad de opinión y expresión, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N14/512/72/PDF/N1451272.pdf?OpenElement
319 Consejo de Derechos Humanos de las Naciones Unidas (2016). Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G16/095/12/PDF/G1609512.pdf?OpenElement
320 Consejo de Derechos Humanos de las Naciones Unidas (2017). Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G17/077/46/PDF/G1707746.pdf?OpenElement
321 Asamblea General de la ONU (2018). Promoción y protección del derecho a la libertad de opinión y expresión, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N18/270/42/PDF/N1827042.pdf?OpenElement
322 Consejo de Derechos Humanos de las Naciones Unidas (2012). Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G12/137/87/PDF/G1213787.pdf?OpenElement7
323 Véase: http://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=21317&LangID=E; DigWatch (2023). Freedom of expression online in 2023, disponible en: https://dig.watch/topics/freedom-expression

Las plataformas digitales están impulsadas por algoritmos que determinan cómo manejar, priorizar, distribuir y eliminar o desechar información de terceros en línea. Existe la posibilidad de que estas actividades no cumplan con los estándares de legalidad, legitimidad y proporcionalidad de las restricciones razonables a la libertad de expresión. Además, las infracciones de la información personal tienen un efecto escalofriante en la libertad de expresión. Las personas se autocensuran y alteran su comportamiento cuando temen ser observadas o carecen de anonimato. Este efecto se verá amplificado por la vigilancia impulsada por la IA, que puede tener un impacto negativo en la libertad de expresión.

Actividad: IA y Libertad de expresión

Los participantes ven el vídeo y debaten las posibles implicaciones de la IA en la libertad de expresión..

Fuente: UNESCO, https://www.youtube.com/watch?v=j0Oz54A68qo

En el mundo digital de hoy, el disfrute de la libertad de expresión se regula en áreas privadas, híbridas y públicas moldeadas por empresas privadas, autoridades gubernamentales y las personas usuarias en relaciones de poder variadas y altamente asimétricas. Además, estos ecosistemas digitales han marcado el camino para nuevos tipos de gobernanza de la expresión, como los moderados por los sistemas de IA en las plataformas de redes sociales para depurar el contenido en las fuentes de noticias de las personas usuarias.

IA, moderación de contenidos y libertad de expresión

Los intermediarios de Internet moderan el contenido en sus plataformas. Esta moderación de contenido a menudo se lleva a cabo fuera de la vista del público y con frecuencia se realiza mediante sistemas de IA opacos a escala, sin garantía de cumplimiento del marco internacional de derechos humanos. Dichos instrumentos automatizados pueden restringir el derecho a la libertad de expresión y el acceso a la información, independientemente del método tecnológico empleado.[324] Pueden excluir del discurso público a individuos, organizaciones, ideas o formas de expresión específicas.

A medida que la cantidad de información en línea que requiere moderación crece inevitable y exponencialmente, las principales plataformas en línea están invirtiendo fuertemente en sistemas de IA para automatizar la moderación del contenido. Un impulso importante para esto es la promulgación de leyes de moderación de contenido en todo el mundo que imponen multas severas por incumplimiento si las plataformas en línea no eliminan rápidamente la información que viola las leyes nacionales de propiedad intelectual, así como las leyes contra el discurso de odio y la pornografía infantil.[325]

Uno de los principales problemas asociados con la automatización de la moderación de contenido es que las tecnologías de IA utilizadas para esto se basan en la tecnología de PNL que es específica del dominio, es decir, la tecnología solo identificará los tipos de contenido en los que se entrenó. Por ejemplo, un sistema de PNL que ha sido entrenado para identificar el discurso racista es incapaz de identificar el contenido violento. Además, incluso dentro de un tema determinado, los algoritmos de PNL podrían no ser capaces de comprender matices detallados del habla humana, como el sarcasmo y la parodia.[326] Un sistema que puede detectar contenido racista en un artículo dentro de un blog puede no reconocer de manera confiable contenido similar en un tweet, lo que resulta en una tasa de error muy alta para estas tecnologías.[327]

> Una guía de utilidad en temas de libertad de expresión y acceso a la información en el entorno digital es **"Salvaguardar la libertad de expresión y el acceso a la información: directrices para un enfoque de múltiples partes interesadas en el contexto de la regulación de las plataformas digitales"** de la UNESCO

Para ilustrar este punto aún más, durante el brote de coronavirus, YouTube reemplazó a muchos de sus revisores de contenido humano con algoritmos de IA encargados de identificar y eliminar videos con desinformación y discurso de odio. El experimento de moderación de contenido en la plataforma falló. Los algoritmos de IA censuraron excesivamente a las personas usuarias, lo que triplicó la tasa de eliminaciones inexactas de contenido. YouTube volvió a contratar a algunos de sus moderadores

324 OSCE (2022). Spotlight on Artificial Intelligence and Freedom of Expression: A Policy Manual, disponible en: https://www.osce.org/representative-on-freedom-of-media/510332

325 Raso F., Hilligoss H., Krishnamurthy V., Bavitz C., Kim L. (2018). Artificial Intelligence & Human Rights: Opportunities & Risks, disponible en: https://cyber.harvard.edu/publication/2018/artificial-intelligence-human-rights

326 Mindmatters (2021). Can the machine know you are just being sarcastic, disponible en: https://mindmatters.ai/2021/05/can-the-machine-know-you-are-just-being-sarcastic/.

327 Ibid. Véase también: Gaumond E., Régis C. (2023). Assessing Impacts of AI on Human Rights: It's Not Solely About Privacy and Nondiscrimination, disponibleen:https://www.lawfareblog.com/assessing-impacts-ai-human-rights-its-not-solely-about-privacy-and-nondiscrimination.

humanos después de unos meses.[328] Otro ejemplo sería el caso de la especialista en moderación de contenido Kate Klonick, que fue expulsada de Twitter por publicar un tweet que contenía la frase "Te asesinaré", que el algoritmo de Twitter consideró como un estímulo a la violencia.[329] Sin embargo, Klonick no estaba incitando a la violencia de ninguna manera. Solo estaba haciendo referencia a un intercambio humorístico entre Molly Jong-Fast y su esposo, quien le iba a quitar la comida.

Vale la pena señalar que las herramientas de PNL aún no son tan efectivas en idiomas diferentes al inglés. Como resultado, las herramientas automatizadas pueden no ser tan precisas en la evaluación de personas que no hablan inglés, lo que puede limitar injustamente su libertad de expresión. Esto es especialmente cierto para las herramientas de traducción, que a veces pueden presentar dificultades con los significados matizados y el contexto. Por ejemplo, ocurrió un incidente en el que un hombre israelí-palestino fue arrestado después de publicar una foto en Facebook con el título "buenos días" en árabe. Sin embargo, la herramienta de traducción impulsada por IA de Facebook tradujo incorrectamente el título a "atacarlos" en hebreo o "lastimarlos" en inglés. Más tarde, Facebook reconoció el error y se disculpó con el hombre y su familia por los inconvenientes causados.[330]

Actividad: Los participantes en la capacitación leen el caso González contra Google y debaten qué leyes serían aplicables en sus jurisdicciones en estas circunstancias. ¿Esto afectaría el resultado del caso?

En 2023, a la Corte Suprema de los Estados Unidos se le presentó un caso interesante, González vs. Google. El caso se planteó después de la trágica muerte de Nohemí González, de 23 años, en los ataques terroristas de París en 2015. La familia de Nohemí González buscó responsabilizar a Google por su papel en los ataques en virtud de la Ley Antiterrorista, que permite a las familias de las personas asesinadas por terroristas emprender acciones legales contra quienes "ayudan e instigan" a tales grupos. Inicialmente, la Corte Suprema se negó a pronunciarse sobre el caso, concretamente sobre si las recomendaciones específicas de los algoritmos de las redes sociales se excluirían de la protección de la Sección 230 de la Ley de Decencia en las Comunicaciones. Esta decisión tiene implicaciones para el futuro de la responsabilidad en casos similares.

Fuente: https://www.supremecourt.gov/opinions/22pdf/21-1333_6j7a.pdf

328 Ibid. Además, véase: Vincent J. (2020). YouTube brings back more human moderators after AI systems over-censor, disponible en: https://www.theverge.com/2020/9/21/21448916/youtube-automated-moderation-ai-machine-learning-increased-errors-takedowns

329 Kloncik K. (2020). What I Learned in Twitter Purgatory, disponible en: https://www.theatlantic.com/ideas/archive/2020/09/what-i-learned-twitter-purgatory/616144/; Gaumond E., Régis C. (2023). Assessing Impacts of AI on Human Rights: It's Not Solely About Privacy and Nondiscrimination, disponible en: https://www.lawfareblog.com/assessing-impacts-ai-human-rights-its-not-solely-about-privacy-and-nondiscrimination

330 Hu X., Neupane B., Flores Echaiz L., Sibal P., Rivera Lam M. (2019). Informe de la UNESCO Steering AI and advanced ICTs for knowledge societies: a Rights, Openness, Access, and Multi-stakeholder Perspective, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000372132

En un entorno en el que las plataformas de redes sociales utilizan algoritmos para decidir qué voces escuchamos, el derecho a la libertad de expresión es de particular importancia. En 2014, investigadores de la Universidad de Cornell realizaron un estudio de contagio emocional en colaboración con Facebook, analizando cómo las emociones se propagan a través de la red social.[331] Los investigadores modificaron las experiencias de más de 700.000 usuarios de Facebook empleando una técnica de análisis de sentimientos para determinar si los amigos habían contribuido con comentarios o publicaciones desagradables. Estos elementos negativos se eliminaron posteriormente de las fuentes de noticias de las personas usuarias en un experimento para determinar si el sesgo algorítmico de la fuente hacia un contenido positivo mantendría a las personas usuarias en el sitio por más tiempo. Este estudio destaca cómo las plataformas pueden tomar decisiones basadas en expresiones de usuario que apoyan una realidad y disminuyen otra.[332]

Actividad: caso de moderación de contenido "La niña del napalm"

El caso de moderación de contenido "La niña del napalm" se refiere a un incidente controvertido que involucra la moderación de imágenes históricas e icónicas relacionadas con la guerra en las plataformas de redes sociales. El caso gira en torno a la eliminación o censura de una fotografía ganadora del Premio Pulitzer conocida como "El terror de la guerra", que muestra a una joven, Kim Phúc, huyendo de un ataque con napalm durante la Guerra de Vietnam. Los participantes de la capacitación leen la descripción general del caso y debaten sus implicaciones para la libertad de expresión en el entorno digital.

Antecedentes:

- La fotografía fue tomada por el fotógrafo de Associated Press (AP) Nick Ut el 8 de junio de 1972, durante la guerra de Vietnam. Captura las consecuencias inmediatas de un bombardeo con napalm en Trang Bang, Vietnam del Sur.
- La imagen muestra a una niña desnuda y gravemente quemada de nueve años, Kim Phúc, corriendo por un camino en agonía.
- La fotografía se ha convertido en un símbolo icónico de los horrores de la guerra y ha desempeñado un papel importante en la sensibilización sobre el costo humano de la guerra de Vietnam.

Incidente de moderación de contenido:

- En septiembre de 2016, Facebook eliminó temporalmente la fotografía cuando fue publicada por el escritor noruego Tom Egeland como parte de una serie de fotografías de guerra icónicas.
- El motivo de la eliminación de Facebook fue su política de no mostrar desnudos en la plataforma.
- La decisión provocó indignación y controversia, y muchos argumentaron que la importancia histórica y periodística de la fotografía debería superar las preocupaciones sobre la desnudez.
- Después de una importante reacción pública y críticas, Facebook revirtió su decisión y restableció la fotografía.

331 Meyer R. (2014). Everything We Know About Facebook's Secret Mood-Manipulation Experiment, disponible en: https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/.

332 Kramer A. D. I., Guillory J. E., Hancock J. T. (2014). Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks, PNAS, 111 (24), disponible en: http://www.pnas.org/content/pnas/111/24/8788.full.pdf

Temas y debates clave:

1. **Libertad de expresión frente a moderación de contenido:** el caso plantea preguntas sobre el equilibrio entre la libertad de expresión, el intercambio de contenido histórico y de interés periodístico, y la necesidad de moderación de contenido para evitar la propagación de material inapropiado u ofensivo.
2. **Moderación algorítmica:** muchas plataformas de redes sociales utilizan algoritmos para detectar y eliminar automáticamente el contenido que infringe sus políticas. En este caso, los algoritmos no lograron distinguir entre una fotografía histórica ganadora del Premio Pulitzer y un contenido inapropiado.
3. **Sensibilidad cultural y contexto:** los críticos argumentan que los algoritmos de moderación de contenido carecen de la capacidad de comprender la importancia histórica, cultural y contextual de ciertas imágenes, lo que lleva a eliminaciones erróneas.
4. **Responsabilidad de las empresas tecnológicas:** el incidente también pone en tela de juicio la responsabilidad de las empresas tecnológicas de tomar decisiones matizadas sobre la moderación de contenido y el impacto potencial de sus decisiones en la libertad de expresión y la documentación histórica.

En última instancia, el caso de moderación de contenido "La niña del napalm" destaca los desafíos que enfrentan las plataformas de redes sociales y las empresas de tecnología para lograr un equilibrio entre moderar el contenido para mantener los estándares de la comunidad y reconocer la importancia del contenido histórico y periodístico, especialmente cuando representa temas sensibles o angustiosos. Subraya la necesidad de políticas y decisiones de moderación de contenido reflexivas y conscientes del contexto.

Fuente: Content Moderation Case Study: Facebook Attracts International Attention When It Removes A Historic Vietnam War Photo Posted By The Editor-in-Chief Of Norway's Biggest Newspaper (2016), disponible en: https://www.techdirt.com/2020/11/20/content-moderation-case-study-facebook-attracts-international-attention-when-it-removes-historic-vietnam-war-photo-posted/

Desinformación e IA

Como ya se señaló, las tecnologías de IA pueden contribuir a un acceso desigual a la información y exacerbar las brechas digitales existentes. Por ejemplo, la IA se puede utilizar para desarrollar y difundir propaganda dirigida, y este problema se ve agravado por los algoritmos de redes sociales impulsados por la IA generados por el "compromiso" que promueven la información en la que es más probable que se haga clic. El análisis de datos utilizado por las empresas de redes sociales para construir perfiles de usuario para publicidad dirigida está impulsado por algoritmos de ML. Además, los bots que se hacen pasar por usuarios genuinos propagan contenido fuera de los grupos de redes sociales fuertemente dirigidos mediante la distribución de enlaces a fuentes falsas y la comunicación activa con las personas como chatbots utilizando el procesamiento del lenguaje natural.[333]

333 Ibid.

Las entidades que implementan algoritmos de evaluación y puntuación de IA con frecuencia no ofrecen una notificación adecuada, si corresponde, a los que se califican y evalúan. Debido a que los consumidores desconocen cómo estas herramientas toman determinaciones y qué tipos de datos emplean, su uso puede erosionar las restricciones relacionadas con el acceso a la información. Debido a que las personas no entienden cómo funcionan estas herramientas, no pueden impugnar las decisiones de elegibilidad que afectan su acceso a servicios, empleos, vivienda o beneficios.[334]

Además, la amenaza de las falsedades, que son sistemas de IA capaces de hacer grabaciones realistas de video y audio de personas reales, ha llevado a muchos a creer que la tecnología se utilizará en el futuro para hacer imágenes falsas de líderes mundiales con fines dañinos. Aunque parece que las falsedades aún no se han utilizado como parte de campañas reales de propaganda o desinformación, y el audio y el video falsificados aún no son convincentemente humanos, la IA detrás de las falsedades está avanzando, y no se debe descartar la posibilidad de propagar el caos, incitar al conflicto y promover la crisis de la verdad.[335]

En las naciones donde la libertad religiosa está amenazada, la IA podría ayudar a los funcionarios gubernamentales a monitorear y atacar a los miembros de las organizaciones religiosas perseguidas. Esto no solo puede aumentar el secreto de tales reuniones por temor a ser detectado, sino que también podría tener consecuencias físicas que van desde el arresto hasta la muerte. Además, la IA podría utilizarse para identificar y eliminar contenido religioso. Si las personas no pueden mostrar símbolos religiosos, orar o enseñar sobre su fe en línea, esto sería una violación flagrante de la libertad religiosa.[336]

La ONG AccessNow señala que el acoso en línea habilitado por bots supone una amenaza clara e inminente para la libertad de expresión. Estas cuentas bot se hacen pasar por usuarios humanos y entregan respuestas automáticas a cuentas designadas o a cualquier persona que comparta un punto de vista particular. Este tipo de acoso en línea implacable tiene un impacto escalofriante en la libertad de expresión, especialmente para los grupos desfavorecidos que son atacados de manera desproporcionada. Los desarrolladores de bots aplican el procesamiento del lenguaje natural con más frecuencia, lo que exacerba las amenazas de acoso en línea por parte de los bots. Esto hará que sea más difícil identificar, informar y eliminar cuentas de bots.[337]

Restricciones legítimas a la libertad de expresión y al acceso a la información

En el marco internacional de derechos humanos y en numerosas constituciones, existen condiciones estrictas para justificar los límites previos a la libertad de expresión y el acceso a la información. En este aspecto, las herramientas de IA son especialmente preocupantes porque estos sistemas están ocultos al escrutinio público, son ciegos al contexto

334 Véase: https://epic.org/issues/ai/screening-scoring/
335 AccessNow (2018). AI and human rights, disponible en: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf
336 Ibid.
337 Ibid.

y funcionan de una manera altamente opaca que impide cualquier corrección o retribución efectiva. Si bien la preselección de contenido para restringir la transmisión en línea de malware y el abuso sexual infantil se ha considerado ampliamente como una aplicación útil de la automatización, se debe tener precaución al aplicar el mismo fundamento a otros tipos de discurso que pertenecen a la categoría más amplia de regulación de contenido.[338] El derecho internacional permite la restricción de los derechos digitales (el derecho a la privacidad, la libertad de expresión y el acceso a la información) tanto fuera de línea como en línea, pero solo en circunstancias muy limitadas y específicas, y de conformidad con el Artículo 19 del Pacto Internacional de Derechos Civiles y Políticos (libertad de expresión y acceso a la información) utilizando la prueba de tres partes que se describe a continuación.[339]

Tabla 6. Prueba de las tres partes para los límites legítimos a la libertad de expresión

Principio	Explicación
Las restricciones deben estar previstas por la ley	• Las leyes de TIC deben estipular claramente cualquier restricción a la libertad de expresión sin ambigüedades. Los ciudadanos deben estar en capacidad de entender y cumplir con las leyes, haciéndolas legítimas. Las disposiciones vagas y demasiado amplias no cumplirían con esta norma. • El Comité de Derechos Humanos de las Naciones Unidas ha declarado en el Comentario General N.º 34 que las restricciones a los derechos digitales deben ser específicas para el contenido. Las prohibiciones generales en ciertos sitios y sistemas no están en línea con el derecho internacional. Además, prohibir la publicación de material basado únicamente en su crítica al gobierno o a su sistema político y social va en contra del derecho internacional.[340]
La restricción debe perseguir un objetivo legítimo	• Según el Artículo (3)19 del Pacto Internacional de Derechos Civiles y Políticos, las limitaciones solo deben imponerse por razones legítimas, como proteger los derechos y la reputación de los demás, garantizar la seguridad nacional, mantener el orden público y promover la salud o la moral públicas.
La restricción debe ser necesaria para un propósito legítimo	• Cualquier limitación al derecho a la libertad de expresión debe ser necesaria y proporcionada. Si bien la vigilancia pública puede ser permisible, los Estados deben demostrar que las medidas son necesarias y proporcionadas. La vigilancia digital es un acto muy intrusivo que viola los derechos digitales. Es necesaria la aprobación previa de una autoridad judicial competente para una vigilancia digital proporcionada. Esto también significa que se utilizarán los métodos de vigilancia menos intrusivos.[341] • Por ejemplo, se ha encontrado que la detección automatizada de amenazas, un sistema comúnmente utilizado por las fuerzas policiales para detectar disparos e identificar posibles escenas del crimen, identifica incorrectamente los sonidos como disparos en el 89% de los casos. Muchos departamentos de policía que anteriormente utilizaban servicios de vigilancia predictiva han descontinuado estos sistemas debido a su utilidad y precisión limitadas.[342]

339 Ibid.
339 Véase: UNESCO (2021). Global Toolkit for Judicial Actors: International legal standards on freedom of expression, access to information and safety of journalists, Module 2, 44–46, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000378755
340 ONU (2011). Comentario general N.º 34, disponible en: https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf
341 International Commission of Jurists, Regulation of Communications Surveillance and Access to Internet in Selected African States, disponible en: https://www.kas.de/documents/275350/0/Report-on-Regulation-of-Communications-Surveillance-and-Access-to-Internet-in-Selected-African-States.pdf/66dbd47d-4d7d-2779-a595-a34e9f93cfbb?t=1639140695434
342 Ibid.

El siguiente video de la UNESCO explica la prueba en tres partes para los límites legítimos a la libertad de expresión:

Fuente: https://www.youtube.com/watch?v=Wg8fVtHPDag

Derecho a la privacidad y a la protección de datos

"Nadie será objeto de injerencias arbitrarias o ilegales en su vida privada, su familia, su domicilio o su correspondencia, ni de ataques ilegales a su honra y reputación. Toda persona tiene derecho a la protección de la ley contra tales injerencias o ataques".

— Artículo 17 del Pacto Internacional de Derechos Civiles y Políticos

La privacidad es fundamental para garantizar otros derechos humanos, incluidos los derechos a la libertad de expresión, opinión, afiliación y reunión. Sin privacidad, a menudo no es práctico ni seguro organizar la oposición política, competir comercialmente o desarrollar alternativas a las políticas existentes, las narrativas dominantes o la injusticia experimentada. La Declaración Universal de Derechos Humanos (DUDH, artículo 12), el Pacto Internacional de Derechos Civiles y Políticos (PIDCP, artículo 17) y varios otros tratados internacionales y regionales de derechos humanos reconocen el derecho a la privacidad como un derecho humano.[343] La importancia del derecho a la privacidad para el ejercicio en línea y fuera de línea de otros derechos humanos, como la libertad de expresión y el acceso a la información, está aumentando en un mundo centrado en los datos.[344]

343 Por ejemplo, la Convención sobre los Derechos del Niño (artículo 16), la Convención Internacional sobre la Protección de los Derechos de Todos los Trabajadores Migratorios y de sus Familiares (artículo 14), la Convención sobre los Derechos de las Personas con Discapacidad (artículo 22), la Carta Africana sobre los Derechos y el Bienestar del Niño (artículo 10), la Convención Americana sobre Derechos Humanos (artículo 11) y el Convenio para la Protección de los Derechos Humanos y de las Libertades Fundamentales (el Convenio Europeo de Derechos Humanos, artículo 8).

344 Consejo de Derechos Humanos de las Naciones Unidas (2021). El derecho a la privacidad en la era digital. The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, disponible en: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

Desde que el PIDCP entró en vigor en 1976, las nuevas tecnologías digitales han evolucionado, y los gobiernos y las organizaciones privadas las han explotado la mayoría de las veces fuera del marco legal y sin tener en cuenta la privacidad individual. Si bien la vigilancia digital y las tecnologías digitales han avanzado rápidamente, la ley de privacidad no ha seguido su ejemplo. Aunque la legislación sobre privacidad a nivel internacional de derechos humanos se basa en principios sólidos y bien establecidos, no ha evolucionado ni se ha modificado para cumplir con los requisitos de la sociedad del siglo XXI. El Comentario General original de 1988 del Comité de Derechos Humanos de la ONU sobre la privacidad no anticipó el desarrollo de nuevas formas de comunicación como el correo electrónico y los mensajes de texto, la aparición de capacidades gubernamentales para interceptar y procesar grandes cantidades de datos electrónicos, o la explosión de sitios web de redes sociales, por nombrar algunos ejemplos.[345]

La Resolución de la Asamblea General de la ONU sobre el derecho a la privacidad en la era digital (2020) se ha referido a "la piratería y el uso ilegal de tecnologías biométricas" como "actos altamente intrusivos que violan el derecho a la privacidad" que interfieren con la libertad de expresión y opinión, la reunión y asociación pacíficas y la libertad religiosa o de creencias, y "pueden contradecir los principios de una sociedad democrática, incluso cuando se realizan de manera extraterritorial o a gran escala".[346] Un informe del Alto Comisionado de las Naciones Unidas para los Derechos Humanos de 2021, "El derecho a la privacidad en la era digital", ha pedido una moratoria en el uso de tecnologías de reconocimiento facial en espacios públicos, hasta que los gobiernos puedan demostrar que no hay problemas sustanciales relacionados con la precisión o los impactos discriminatorios y que estas tecnologías cumplen con estándares sólidos de privacidad y protección de datos.[347]

Privacidad y protección de datos en el ámbito digital

Comprender la ley de protección de datos y privacidad en el ámbito digital requiere una comprensión integral de la definición, clasificación y aparición de la privacidad como una preocupación social. El derecho a la privacidad es un principio fundamental para una sociedad democrática y desempeña un papel crucial en el equilibrio de poder entre el gobierno, las entidades del sector privado que recopilan, procesan y almacenan datos personales, y las personas cuyos datos personales se recopilan, procesan y almacenan. La importancia del derecho a la privacidad para el ejercicio en línea y fuera de línea de otros derechos humanos, como la libertad de expresión y el acceso a la información, está aumentando en un mundo centrado en los datos.[348]

345 American Civil Liberties Union (2015). Information Privacy in the Digital Age, disponible en: https://www.aclu.org/other/human-right-privacy-digital-age

346 Consejo de Derechos Humanos de las Naciones Unidas (2021). El derecho a la privacidad en la era digital. The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, disponible en: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

347 Ibid.

348 Consejo de Derechos Humanos de las Naciones Unidas (2021). El derecho a la privacidad en la era digital. The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, disponible en: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

Desde el inicio de la era de la información, el derecho a la privacidad y la necesidad de proteger la información o los datos personales han recibido una atención considerable. Vivimos en una era en que las tecnologías digitales permiten la recopilación, el almacenamiento y el procesamiento masivos y rentables de datos personales en línea, así como el monitoreo de las personas dondequiera que se encuentren (incluido el monitoreo de sus actividades en línea). Si bien Internet y el intercambio de información en línea y la recopilación de datos aumentan a un ritmo exponencial, los desarrollos legislativos no han logrado mantener el ritmo y proteger adecuadamente la información personal. Los gobiernos de todo el mundo han comenzado a adoptar instrumentos y regulaciones relacionados con la protección de datos para proteger los derechos de privacidad de sus ciudadanos.[349]

El concepto de privacidad se compone de un conjunto de principios. El derecho a la privacidad garantiza que un espacio está reservado para la autoexpresión. De esta manera, el derecho está fuertemente relacionado con la libertad de expresión. Cada vez se reconoce más que el derecho a la privacidad desempeña un papel vital para facilitar el derecho a la libertad de expresión y el acceso a la información. Por ejemplo, la protección del derecho a la privacidad permite a las personas compartir opiniones de forma anónima en circunstancias en las que pueden temer ser censuradas por esas opiniones, permite a los denunciantes hacer divulgaciones protegidas y permite a los miembros de los medios y activistas comunicarse de forma segura más allá del alcance de la vigilancia gubernamental.[350] Además, el derecho a la privacidad salvaguarda la intimidad y la dignidad. Además, incluye el derecho a decidir cómo vivir y el derecho a la autonomía en general. El derecho a la privacidad incluye la privacidad informativa, así como el derecho a acceder y controlar la información personal, independientemente de su formato. Estos subcomponentes de privacidad no son exhaustivos; más bien, sirven como una hoja de ruta para el futuro desarrollo de medidas de privacidad en el entorno digital.[351]

La línea entre el mundo en línea y fuera de línea se está volviendo cada vez más difusa. De hecho, parece que las personas viven en un estado continuo de conexión y desconexión, lo que dificulta la definición de límites claros. Con la ayuda de la IA, las organizaciones (tanto privadas como gubernamentales) pueden recopilar, procesar y reutilizar fácilmente grandes cantidades de datos e imágenes, que incluyen datos confidenciales de las personas usuarias. Los algoritmos de IA permiten hacer predicciones sobre la vida personal de la gente, como sus hábitos de sueño e incluso su lugar de residencia.

Las empresas de redes sociales prosperan en la recopilación y comercialización de grandes volúmenes de datos de usuarios de Internet, lo que enfatiza aún más la necesidad de proteger la privacidad de la persona usuaria en el mundo en línea y fuera de línea. De hecho, “la gente parece vivir en un continuo estado de conexión/desconexión, con el resultado de que es difícil trazar líneas nítidas y significativas entre los dos”. La IA facilita

349 Defensa de los medios (2022). Módulo 4: Data Privacy and Data Protection, disponible en: https://www.mediadefence.org/ereader/publications/modules-on-litigating-freedom-of-expression-and-digital-rights-in-south-and-southeast-asia/module-4-data-privacy-and-data-protection/introduction/
350 Ibid.
351 American Civil Liberties Union (2015). Information Privacy in the Digital Age, disponible en: https://www.aclu.org/other/human-right-privacy-digital-age

la recopilación, el procesamiento y la reutilización de cantidades masivas de datos e imágenes, alentando a las organizaciones (tanto del sector privado como del gobierno) a recopilar, retener y manejar datos confidenciales sobre las personas usuarias. Los algoritmos de IA hacen predicciones sobre la vida personal de las personas, incluyendo cosas como dónde viven y sus hábitos de sueño.

A medida que avanzamos en nuestra vida diaria, los rastreadores GPS de nuestros teléfonos inteligentes pueden recopilar una gran cantidad de datos sobre nuestros movimientos, incluso si no estamos usando Internet activamente. Cuando visitamos lugares como cafeterías, escuelas e instalaciones médicas, esta información se puede utilizar para hacer inferencias sobre nuestra identidad personal, intereses, aspiraciones, problemas y redes sociales en función de cuánto tiempo permanecemos allí y los movimientos de los demás a nuestro alrededor. Estos datos pueden ser bastante reveladores y pueden tener implicaciones significativas para nuestra privacidad y seguridad. Por ejemplo, cuando nos movemos por la ciudad y vamos a una cafetería, una escuela o una institución médica, el rastreador GPS en nuestros teléfonos inteligentes puede detectar dónde estamos y cuánto tiempo permanecemos allí y recopilar estos datos (y correlacionarlos con los movimientos de otros), incluso si no accedimos a Internet en nuestros teléfonos. Se pueden derivar inferencias significativas con respecto a nuestra identidad, intereses, aspiraciones, problemas y redes a partir de dichos datos.

Las formas nuevas y económicas de análisis y almacenamiento de datos, junto con la conectividad digital y en línea mejorada (desde electrodomésticos inteligentes hasta nanobots dentro de los cuerpos humanos) y las tecnologías emergentes como la IA y el IoT han permitido a los gobiernos y corporaciones gigantes convertirse en mineros de datos, recopilando información sobre todos los aspectos de las actividades humanas, el comportamiento y el estilo de vida.

La normativa de privacidad se ha adaptado a los novedosos retos que plantea el entorno digital y en línea. Muchas naciones de todo el mundo han implementado regulaciones que requieren el consentimiento de los interesados para usar y procesar sus datos personales en línea, garantizar el acceso a los datos personales por parte de los interesados y otorgar el derecho a que estos datos personales sean eliminados, corregidos o transferidos a una entidad diferente.

Las leyes que preservan la privacidad en el entorno de la IA tienen como objetivo equipar a las personas con el derecho a ver el contenido de las bases de datos que contienen información sobre ellas. Estas leyes también tienen como objetivo restringir el uso de información personal sin el consentimiento del interesado, excepto en circunstancias limitadas definidas por la ley. Según estas leyes, las personas tienen derecho a aceptar los términos de uso antes de descargar una aplicación en su teléfono celular o comenzar a usar software gratuito, es decir, productos y servicios cuyo modelo económico se basa en la comercialización de datos personales.[352]

352 Altshuler TS (2019). Privacy in a digital world, disponible en: https://techcrunch.com/2019/09/26/privacy-queen-of-human-rights-in-a-digital-world

Los datos personales, que se almacenan en línea, a menudo se procesan de numerosas maneras y propósitos, algunos de los cuales no se pueden anticipar en el momento en que el interesado otorga el consentimiento. Además, muchos de nosotros rara vez pasamos por los términos de uso, incluso cuando son concisos y se muestran en letra grande.[353] Por ejemplo, nos llevará 76 días leer las políticas de privacidad que uno puede encontrar cada año.[354]

Otro aspecto de la privacidad en el entorno de la IA es entender la privacidad como el "derecho a estar en paz".[355] Esto se refiere al derecho a mantener un espacio seguro y protegido alrededor de nuestro cuerpo, pensamientos íntimos, sentimientos y estilo de vida cuando estamos en línea. El monitoreo constante en línea de nuestras acciones mediante sensores, cámaras de vigilancia, asistentes digitales, como Siri, Alexa y otras herramientas digitales y de inteligencia artificial, puede tener un profundo impacto en el derecho a la privacidad como un derecho humano.[356]

Caso de estudio: Grabación y envío de conversaciones privadas de Amazon Alexa

Una familia en Oregón, EE. UU., informó que su dispositivo Amazon Echo había grabado una conversación privada que estaban teniendo en su casa. Aún más preocupante, la conversación grabada fue enviada a un contacto en la libreta de direcciones de la familia, un colega de uno de los miembros de la familia, sin su consentimiento o conocimiento. El incidente salió a la luz cuando el destinatario de la conversación grabada se puso en contacto con la familia para informarles sobre el mensaje inusual. Amazon investigó el incidente y lo atribuyó a una combinación extremadamente rara de circunstancias. Según Amazon, el dispositivo Echo había interpretado erróneamente partes de la conversación como comandos para enviar un mensaje. Fue un caso de "falsos positivos" de detección de palabras de activación, donde el dispositivo pensó erróneamente que había escuchado la palabra de activación (probablemente "Alexa") y comenzó a grabar y enviar la conversación. Amazon consideró en serio el incidente y tomó medidas para mejorar la tecnología de reconocimiento de palabras de activación para evitar tales falsos positivos. La compañía también introdujo una función que permite a las personas usuarias agregar un PIN a las compras por voz para evitar pedidos accidentales a través de comandos de voz. Este incidente provocó debates sobre la privacidad y la seguridad de los dispositivos activados por voz, lo que llevó a una mayor conciencia y preocupación de las personas usuarias sobre la posibilidad de escuchas. En general, el incidente destacó la necesidad de que las empresas de tecnología mejoren continuamente las características de privacidad y seguridad de los dispositivos activados por voz como Amazon Alexa. También enfatizó la importancia de la educación de la persona usuaria con respecto a la configuración del dispositivo y los controles de privacidad para garantizar una experiencia de usuario más segura.

Fuente: Wolfson S. (2018). Amazon's Alexa recorded private conversation and sent it to random contact, available at: https://www.theguardian.com/technology/2018/may/24/amazon-alexa-recorded-conversation

353 Ibid.
354 Popkin H. A. S. (2012). Life is too short to read privacy policies - here 's statistical proof!, disponible en: https://www.nbcnews.com/tech/tech-news/life-too-short-read-privacy-policies-heres-statistical-proof-flna297399
355 Altshuler TS (2019). Privacy in a digital world, disponible en: https://techcrunch.com/2019/09/26/privacy-queen-of-human-rights-in-a-digital-world
356 Ibid.

Es importante tener en cuenta que las empresas de tecnología han tomado medidas para abordar estas preocupaciones y mejorar la privacidad de la persona usuaria al proporcionar más transparencia, mejorar la configuración de privacidad y permitir que las personas usuarias eliminen las grabaciones de voz. Sin embargo, estos incidentes resaltan la necesidad de que las personas usuarias estén atentos a su configuración de privacidad y a los riesgos potenciales asociados con los dispositivos activados por voz. Las personas usuarias también deben conocer las prácticas de recopilación y almacenamiento de datos de los asistentes virtuales que utilizan y tomar decisiones informadas sobre su uso.

Perfiles de IA

Un tercer aspecto de la privacidad en el entorno de la IA es el derecho a oponerse a la creación automática de perfiles al limitar la capacidad de las entidades comerciales o gubernamentales para combinar datos personales con big data acumulados de otras personas para construir perfiles de comportamiento utilizando la IA y el aprendizaje automático.[357] Las herramientas de IA se utilizan para buscar patrones en el comportamiento humano. Tener acceso a los conjuntos de datos correctos puede utilizarse para hacer inferencias sobre cosas cotidianas que son profundamente privadas y personales, como cuántos residentes de un vecindario es probable que visiten un lugar de culto específico, qué programas de televisión podrían disfrutar e incluso aproximadamente sus patrones de sueño.

El uso de técnicas de IA puede identificar grupos, como aquellos que comparten una postura política o personal específica, y sacar conclusiones amplias sobre las personas, incluso sobre su salud mental y física. A pesar de su carácter probabilístico, los juicios y predicciones proporcionados por la IA a menudo pueden servir como base para decisiones que tienen un impacto en los derechos fundamentales de las personas. Estos problemas se agravan en el contexto del poder judicial, por ejemplo, cuando los jueces confían en la toma de decisiones con la ayuda de los sistemas de IA.[358]

La historia de cómo Target utilizó el análisis de datos para predecir que una adolescente estaba embarazada antes de que su familia lo supiera es un ejemplo bien conocido del poder del análisis de datos y el modelado predictivo en el comercio minorista. Aquí se presenta un resumen del caso:

En 2012, un artículo en The New York Times reveló que Target, un gigante minorista estadounidense, había desarrollado un algoritmo para predecir los hábitos y preferencias de compra de los clientes. Utilizaron estos datos para enviar anuncios y cupones dirigidos a los clientes. Uno de los ejemplos más famosos de este artículo involucró a una adolescente.

El algoritmo de Target había identificado que una adolescente estaba comprando loción sin perfume, suplementos dietéticos y bolas de algodón. Si bien estas compras pueden parecer no relacionadas, el algoritmo reconoció que esta combinación de productos a menudo era indicativa de embarazo. El algoritmo asignó una puntuación

357 Ibid.

358 Consejo de Derechos Humanos de las Naciones Unidas (2021). El derecho a la privacidad en la era digital. The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, disponible en: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx

de “predicción de embarazo” a cada cliente en función de su historial de compras. Una vez que el sistema tuviera una puntuación de predicción alta para un cliente, comenzaría a enviarle anuncios y cupones relacionados con productos para el embarazo y el bebé. En este caso específico, Target comenzó a enviar a la adolescente cupones para productos para bebés como pañales, cunas y ropa de bebé.

El padre de la niña se sorprendió al encontrar estos anuncios relacionados con el embarazo dirigidos a su hija. Llamó a la tienda para quejarse de los anuncios inapropiados que pensaba que le estaban enviando a su hija adolescente. Sin embargo, unos días más tarde, descubrió que su hija estaba embarazada.

El algoritmo había predicho con precisión el embarazo de la chica en función de sus patrones de compra, incluso antes de que su familia lo supiera. La combinación de productos aparentemente no relacionados en su historial de compras, como lociones sin perfume y bolas de algodón, indicaba una alta probabilidad de embarazo.

Este caso ilustra cómo los minoristas pueden utilizar el análisis avanzado de datos y el modelado predictivo para comprender el comportamiento de los clientes y enviar anuncios altamente orientados. Si bien esto puede ser efectivo para fines de marketing, también plantea preguntas importantes sobre la privacidad y la ética de la recopilación y el uso de los datos de los clientes. Es esencial que las empresas manejen los datos de los clientes de manera responsable y transparente para mantener la confianza de sus clientes.

Fuente: Duhigg C. (2012). How Companies Learn Your Secrets, disponible en: https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Para conocer los hechos del caso Nubian Rights Forum y otros vs. The Attorney General, Kenia, 2021 y para analizar sus implicaciones en la vigilancia digital y la privacidad en Kenia, lea el artículo de Privacy International **“Data Protection Impact Assessments and ID systems: the 2021 Kenyan ruling on Huduma Namba”**

Las herramientas de IA también se pueden utilizar para perfilar a los jueces. Una iniciativa regulatoria interesante que tiene como objetivo salvaguardar la integridad del juez y evitar la elaboración de perfiles por parte de la IA es la ley francesa sobre Programación y Reforma de la Justicia (2019-222). En su artículo 33, esta regulación tiene como objetivo evitar que cualquier persona, pero especialmente las empresas de tecnología legal enfocadas en predecir y analizar litigios, divulguen públicamente el patrón de comportamiento de los jueces en relación con las decisiones judiciales. Dice lo siguiente: “Los datos de identidad de los jueces y miembros del poder judicial no podrán ser reutilizados con el propósito o efecto de evaluar, analizar, comparar o predecir sus prácticas profesionales reales o supuestas”.[359]

Muchos gobiernos han comenzado a digitalizar sus servicios públicos, poniéndolos en línea y ofreciendo sistemas nacionales de identificación digital (ID). Al acumular grandes volúmenes de datos personales, estos sistemas digitales y bases de datos amenazan el derecho a la privacidad de los ciudadanos. Los programas nacionales de identidad digital son solo uno de los muchos ejemplos de cómo los gobiernos pueden violar los derechos digitales. Estos programas requieren recopilar y almacenar datos personales

359 Ley francesa de Programación y Reforma de la Justicia (2019-222), disponible en: https://www.wipo.int/wipolex/en/legislation/details/18789

confidenciales e identificadores biométricos para crear una sola identificación digital, a fin de mejorar la prestación de servicios gubernamentales. Sin embargo, es importante que los gobiernos comprendan los riesgos potenciales para las personas usuarias antes de crear bases de datos centralizadas de datos personales y biométricos. Para prevenir las violaciones de los derechos humanos y la ciberseguridad, las leyes deben incluir las protecciones adecuadas antes de implementar dichos programas. Muchos tribunales nacionales y regionales han actuado en demandas contra estos sistemas digitales presentadas por ciudadanos y organizaciones de la sociedad civil (OSC).

Uno de esos casos es Nubian Rights Forum y otros vs. El Fiscal general, Kenia, 2021, donde el Tribunal Superior de Kenia declaró inconstitucional el Sistema Nacional Integrado de Gestión de Identidad (NIIMS) del país, un sistema de identificación digital.[360] El Tribunal declaró que una Evaluación de Impacto de Protección de Datos debería haber precedido al programa y que debería haber existido un marco legal adecuado para mitigar los riesgos de privacidad y protección de datos antes de la implementación del NIIMS.[361] Este pasaje destaca los escollos comunes que las sentencias judiciales en varios países han identificado al decidir los desafíos de las OSC y otras partes interesadas contra los sistemas de identificación digital. En otro caso, la Corte Suprema de Mauricio enfatizó la falta de una defensa adecuada contra los riesgos de seguridad asociados con la biometría. La sentencia de Aadhaar en la India expresó su preocupación por las bases de datos centralizadas, mientras que el Tribunal Supremo de Filipinas identificó el riesgo de seguimiento individual a través de un sistema de identidad nacional. Finalmente, el Tribunal Superior de Kenia identificó el riesgo de exclusión debido a fallas en el registro biométrico y otros sistemas de identidad.[362]

Actividad: Publicidad dirigida y discriminación de precios impulsada por algoritmos de IA. Los participantes en la capacitación analizan los principales problemas legales y de derechos humanos afectados por la publicidad dirigida y los precios personalizados. ¿Qué leyes son aplicables en estas circunstancias?

Publicidad dirigida

En la era digital actual, los algoritmos de autoaprendizaje se han convertido en una parte integral del análisis de big data. Con la ayuda de la IA, las empresas privadas pueden recopilar una gran cantidad de información personal, como sus hábitos de navegación, los me gusta de las redes sociales, los datos de salud y los patrones de compra. Estos datos se pueden utilizar para crear un perfil detallado de una persona, que se puede utilizar aún más para el seguimiento y la elaboración de perfiles en línea. Esto ayuda a las empresas a adaptar su publicidad, precios y términos contractuales al perfil específico del cliente, y aprovechar los sesgos y la voluntad de pago del consumidor, todo gracias a los hallazgos de la economía del comportamiento. Además, los conocimientos basados en IA también se pueden utilizar para los sistemas de puntuación, que pueden decidir si un consumidor específico cumple los requisitos para comprar un producto o contratar un servicio en particular. El uso de algoritmos de autoaprendizaje en el análisis de big data permite a las empresas privadas obtener una visión detallada de las circunstancias personales, los patrones de comportamiento y la

360 Véase: https://globalfreedomofexpression.columbia.edu/cases/nubian-rights-forum-v-attorney-general.

361 © UNESCO 2022 Directrices para los actores judiciales sobre privacidad y protección de datos, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000381298

362 Privacy International (2022). Data Protection Impact Assessments and ID systems: the 2021 Kenyan ruling on Huduma Namba, disponible en: https://privacyinternational.org/news-analysis/4778/data-protection-impact-assessments-and-id-systems-2021-kenyan-ruling-huduma

personalidad (compras, sitios visitados, me gusta en las redes sociales, datos de salud). La IA se utiliza en el seguimiento y la elaboración de perfiles en línea de personas cuyos hábitos de navegación se recopilan mediante "cookies" y huellas digitales y luego se combinan con consultas a través de motores de búsqueda o asistentes virtuales. Las empresas pueden adaptar su publicidad, precios y términos contractuales al perfil del cliente respectivo y, basándose en los hallazgos de la economía del comportamiento, explotar los sesgos del consumidor y/o su disposición a pagar. Los conocimientos basados en IA también se pueden utilizar para que los sistemas de puntuación decidan si un consumidor específico puede comprar un producto o contratar un servicio.

El creciente uso de publicidad dirigida, que se basa en el seguimiento y la elaboración de perfiles en Internet, ha suscitado preocupaciones sobre la privacidad y la protección de datos. Con todo automatizado, las personas usuarias a menudo no pueden dar un consentimiento significativo. El uso de IA para el procesamiento intensivo de datos puede exacerbar aún más otras violaciones de derechos, particularmente en los casos en que los datos personales se utilizan para objetivizar a personas en contextos como las solicitudes de seguro o empleo. En algunos casos, los algoritmos pueden incluso representar una amenaza tanto para el derecho a la privacidad como para la libertad de expresión. Esto crea problemas crecientes para la privacidad y la protección de datos. La publicidad dirigida utiliza el seguimiento y la creación de perfiles en Internet en función de los intereses esperados de la persona. Todos estos métodos han incapacitado a las personas usuarias para dar un consentimiento significativo porque todo está automatizado. El procesamiento intensivo de datos mediante IA puede exacerbar otras violaciones de derechos cuando los datos personales se utilizan para dirigirse a personas, como en el contexto de solicitudes de seguro o empleo, o cuando los algoritmos amenazan tanto el derecho a la privacidad como la libertad de expresión.[363] Por ejemplo, los algoritmos de las redes sociales deciden el contenido del suministro de noticias de un usuario e influyen en el número de personas que ven y comparten información. Los algoritmos de los motores de búsqueda indexan el contenido y determinan lo que aparece en la parte superior de los resultados de búsqueda. Estos algoritmos amenazan el pluralismo de los medios y suprimen la diversidad de puntos de vista.[364] Para ilustrar el asunto, en 2023, el Comité de Protección de Datos de Irlanda multó a Meta con 390 millones de euros por violar el RGPD. El regulador ha alegado que el uso de datos personales de Meta en Facebook e Instagram, específicamente para publicidad personalizada, no cumplía con el RGPD.[365]

Discriminación de precios

En la era digital, la IA desempeña un papel importante para ayudar a las empresas a adaptar sus ofertas a los clientes individuales. Al analizar el comportamiento y las preferencias de los consumidores, los algoritmos de IA pueden estimar el precio más alto que un cliente en particular está dispuesto o puede pagar. Este enfoque es particularmente relevante para industrias como el crédito y los seguros, que operan en estructuras de costos basadas en el riesgo que tienen en cuenta las características únicas de cada consumidor. Sin embargo, la cuestión de si los reguladores deberían permitir la discriminación de precios en otros sectores en función de la capacidad de pago de un cliente es un tema complejo y polémico que requiere una mayor exploración y debate. La IA apoya a las empresas digitales para que presenten a los consumidores precios individualizados y ofrezcan a cada consumidor una aproximación del precio más alto que el consumidor puede o puede estar dispuesto a pagar. Ciertos mercados, como el crédito o los seguros, operan en estructuras de costos basadas en perfiles de riesgo correlacionados con

363 Consejo de Europa (2017). Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications, disponible en: https://rm.coe.int/study-hr-dimension-of-automated-data-processing-incl-algorithms/168075b94a

364 Access Now (2018). Human rights in the age of artificial intelligence, disponible en: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

365 La Comisión de Protección de Datos (2023). Data Protection Commission announces conclusion of two inquiries into Meta Ireland, disponible en: https://www.dataprotection.ie/en/news-media/data-protection-commission-announces-conclusion-two-inquiries-meta-ireland

características distintivas de los consumidores individuales, lo que sugiere que puede ser razonable ofrecer diferentes precios (por ejemplo, tasas de interés) a diferentes consumidores. ¿Deberían los reguladores permitir la discriminación de precios también en otros casos, en función de la capacidad de pago de los diferentes consumidores?[366]

Es preocupante que los consumidores generalmente no sepan cuándo se han personalizado la publicidad, la información, los precios o los términos del contrato en función de su perfil. Si un algoritmo calcula una determinada puntuación que da como resultado que un contrato no se ofrezca o solo se ofrezca en condiciones desfavorables, los consumidores a menudo tienen dificultades para comprender cómo se generó esta puntuación. Además, la complejidad, la imprevisibilidad y el comportamiento semiautónomo de los sistemas de IA pueden plantear desafíos para hacer cumplir la legislación de los consumidores, ya que es difícil rastrear las decisiones hasta un solo actor y garantizar el cumplimiento legal. Los consumidores generalmente no son conscientes de que la publicidad, la información, los precios o los términos del contrato se han personalizado de acuerdo con su perfil. Supongamos que un determinado contrato no se concreta o solo se ofrece en condiciones desfavorables debido a una determinada puntuación calculada por un algoritmo. En ese caso, los consumidores a menudo no pueden entender cómo se logró esta puntuación. La complejidad, la imprevisibilidad y el comportamiento semiautónomo de los sistemas de IA también pueden dificultar la aplicación efectiva de la legislación del consumidor, ya que la decisión no se puede rastrear a un actor singular y, por lo tanto, no se puede verificar el cumplimiento legal.

Todas estas prácticas de elaboración automatizada de perfiles habilitadas por la IA han tenido graves implicaciones para el disfrute del derecho a la vida privada y familiar. Los rastros de información personal, como el escape digital producido a sabiendas o sin saberlo por teléfonos celulares, computadores y otras tecnologías, que quedan en el ámbito digital son interminables. La forma en que esa información personal es recopilada y utilizada por terceros es una gran preocupación para los reguladores.[367]

La IA se utiliza en el seguimiento y la elaboración de perfiles en línea de personas cuyos hábitos de navegación se recopilan mediante “cookies” y huellas digitales y luego se combinan con consultas a través de motores de búsqueda o asistentes virtuales. Las aplicaciones móviles procesan datos de comportamiento (como datos de ubicación y salud) de dispositivos inteligentes. Esto crea problemas crecientes para la privacidad y la protección de datos. La publicidad dirigida utiliza el seguimiento y la creación de perfiles en Internet en función de los intereses esperados de la persona. El uso

Para obtener una experiencia de primera mano del seguimiento en línea, los participantes de la capacitación deben conectarse al administrador de preferencias de anuncios de Google en: http://www.google.com/ads/preferences/ y mirar los marcadores utilizados por la empresa para definirlos y evaluar qué tan precisos son. La información rastreada se utiliza para crear perfiles digitales de los usuarios a los que se vende el acceso en el mercado, incluidos los intercambios especializados, para ayudar a los anunciantes a comercializar mejor sus productos.

366 Parlamento Europeo (2019). Artificial Intelligence: Challenges for EU Citizens and Consumers, disponible en: https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/631043/IPOL_BRI(2019)631043_EN.pdf

367 Perry W. L., McInnis B., Price C. C., Smith S., Hollywood J. S. (2013). Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations, RAND Corporation: Santa Monica, disponible en: https://www.rand.org/pubs/research_reports/RR233.html

de todos estos métodos ha incapacitado a las personas usuarias para dar un consentimiento significativo porque todo está automatizado. A pesar de que se puede solicitar el consentimiento de las personas usuarias según lo exige la ley (a) no siempre entienden necesariamente lo que se les pide; (b) aun así, la terminología y los términos y condiciones pueden ser confusos y encontrarse en muchas páginas; y (c) con tanto contenido en línea, las personas usuarias sufren de sobrecarga de información.

Caso de estudio: jurisprudencia sobre la elaboración de perfiles de personas a través de ADM

En 2018, la Autoridad Italiana de Protección de Datos (Garante) descubrió que un controlador de datos estaba violando la ley nacional de protección de datos al ofrecer tarifas personalizadas a los clientes de su servicio de uso compartido de automóviles en función de sus hábitos y características observados. En el procedimiento administrativo, la demandada se defendió, alegando que no existía una "categorización" de las personas usuarias del servicio porque la información utilizada para determinar las tarifas no estaba vinculada a los sujetos. La DPA rechazó las objeciones de la demandada, encontrando que era evidente la existencia de un tratamiento de datos personales en este caso, que se trataba exclusivamente de un tratamiento automatizado y que tenía como objetivo definir el perfil o la personalidad de una persona o analizar sus hábitos o elecciones de consumo. La Corte Suprema de Italia (Corte Suprema di Cassazione) confirmó esta decisión en noviembre de 2021, lo que resultó en una multa administrativa de 60.000 euros. En el proceso de apelación, la Corte Suprema se puso del lado del Garante porque dictaminó que el procesamiento de datos personales utilizando un algoritmo para determinar una tarifa individual constituye un perfil, incluso si los datos no son almacenados por el controlador ni atribuibles al interesado.

Fuente: Future of Privacy Forum (2022). GDPR and the AI Act interplay: Lessons from FPF's ADM Case Law Report, disponible en: https://fpf.org/blog/gdpr-and-the-ai-act-interplay-lessons-from-fpfs-adm-case-law-report

La anonimización de datos no siempre conduce a la protección de la privacidad

The privacy of data is usually protected through anonymization. Identifiable aspects such as names, phone numbers, and email addresses are stripped out. Datasets are altered to be less precise, and "noise" is introduced to the data. However, a study published by Nature Communications suggests that anonymization does not always protect privacy. Researchers have developed an ML model that estimates how individuals can be re-identified from an anonymized data set by entering their zip code, gender, and date of birth.

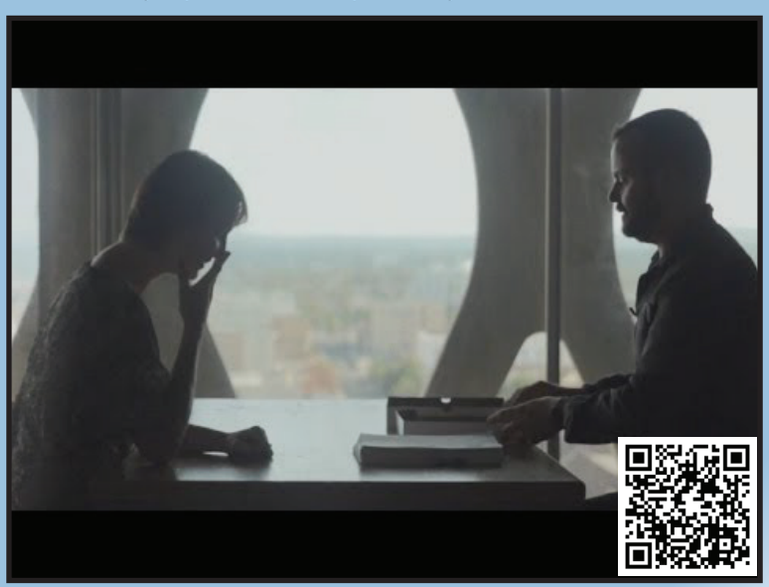
Fuente: Rocher L., Hendrickx J. M., de Montjoye Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models, Nature Communications, 10 (3069), available at: https://www.nature.com/articles/s41467-019-10933-3

Problemas de privacidad emergentes

La creación de nuevos datos es un reto único en el tratamiento automatizado de datos personales. A menudo es posible combinar datos personales, lo que lleva a la creación de segundas e incluso terceras generaciones de datos sobre una persona en particular. En comparación con un conjunto de datos mucho más grande, dos piezas de información aparentemente no relacionadas podrían "reproducirse" y crear nuevos datos, sin el conocimiento del interesado. Se plantean preguntas significativas sobre los conceptos de consentimiento, apertura y autonomía personal.[368] Cuestiones que merecen mayor atención: ¿cuánto control tendrán los sujetos sobre la información recopilada sobre ellos? Dada su participación en la provisión de datos personales para fines de capacitación en ML, ¿deberían las personas tener derecho a utilizar el modelo o al menos saber para qué se utiliza? ¿Podrían los sistemas de ML de búsqueda de datos violar inadvertidamente la privacidad de las personas si, por ejemplo, el análisis del genoma de un miembro de la familia revelara datos de salud sobre otros miembros de la familia?[369]

Punto de debate (10 a 15 minutos): "El poder de la privacidad (1/5): ¿Internet sabe dónde vive?"

Los participantes de la capacitación ven el video producido por The Guardian, "El poder de la privacidad (1/5): ¿Internet sabe dónde vive?" y debaten cómo ha cambiado la noción de privacidad en el ámbito digital y cómo esto ha impactado su trabajo. También debaten ejemplos de sus respectivas jurisdicciones.



Fuente: https://www.youtube.com/watch?v=iA89GhyLao8

368 Comité de expertos en intermediarios de Internet (MSI-NET) (2018). Algorithms and Human Rights: Study on the Human Rights Dimensions of Automated Data Processing Techniques and Possible Regulatory Implications, Council of Europe Study, DGI/2017/12, disponible en: https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5

369 Parlamento Europeo (2020). La ética de la inteligencia artificial: problemas e iniciativas, disponible en: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)634452

Todos estos desafíos se han exacerbado en entornos del sector público. Según la ONG Access Now, con la expansión de Internet y el crecimiento de las nuevas tecnologías, la vigilancia gubernamental ha aumentado y la IA está permitiendo capacidades de vigilancia más intrusivas que nunca. Aunque actualmente no se conoce ningún sistema de reconocimiento facial gubernamental completamente centralizado, algunos países han intentado desplegar más cámaras de CCTV en áreas públicas y centralizar sus sistemas de reconocimiento facial.[370] La mitad de todos los adultos estadounidenses se encuentran ahora en las bases de datos de reconocimiento facial de las fuerzas del orden.[371] El uso de estas tecnologías supone una amenaza para el anonimato, y el temor a ser observado puede impedir el ejercicio de otros derechos, como la libertad de asociación. Los grupos demográficos desfavorecidos, que ya están bajo el control frecuente de las fuerzas de seguridad, experimentarían los efectos negativos de la vigilancia impulsada por la IA de manera más directa. Además, dado que monitorear a toda la población las 24 horas del día, los siete días de la semana no es esencial ni proporcional al propósito de la seguridad pública o la prevención del delito, es casi probable que viole el derecho a la privacidad.[372]

370 AI and human rights, disponible en: https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf
371 Ibid.
372 Ibid.

Casos de estudio: El caso del sistema de reconocimiento facial SARI Real Time, Italia

La autoridad italiana de protección de datos (DPA) ha publicado una opinión sobre el sistema Sari Real Time presentado para su revisión por el Ministerio del Interior del país, manifestando que si se emplea según lo previsto, la tecnología "establecería un tipo de monitoreo masivo". Sari, que aún no está operativo, es un sistema de reconocimiento facial que, utilizando varias cámaras ubicadas en áreas geográficas específicas, analizaría los rostros de las personas filmadas en tiempo real y los compararía con una base de datos preparada de hasta 10.000 rostros. El Sari se implementaría "cuando exista la necesidad de una tecnología de reconocimiento facial para ayudar a las fuerzas policiales en la gestión del orden y la seguridad pública, o en respuesta a los requisitos de la policía judicial".

El DPA ha declarado que Sari 'llevaría a cabo un procesamiento automatizado a gran escala que podría incluir a aquellos presentes en manifestaciones políticas y sociales que no son objeto de "atención" policial'. Además, el hecho de que "la identificación de una persona se lograría mediante el procesamiento de los datos biométricos de todas las personas presentes en el espacio monitoreado" resultaría en una "transición de la vigilancia selectiva de individuos específicos a la perspectiva de una vigilancia universal". El DPA determinó que el Ministerio no había aclarado el sustento legal sobre el cual llevaría a cabo tales acciones. Se afirmó que "un marco regulatorio efectivo debe tener en cuenta todos los derechos y libertades involucrados e identificar los escenarios en los que el uso de dichos sistemas es permisible, sin dar mucha discreción a las personas usuarias".

Fuente: DigWatch, autoridad italiana de protección de datos: El sistema de reconocimiento facial Sari propuesto por el Ministerio del Interior podría dar lugar a una vigilancia masiva, disponible en: https://dig.watch/updates/italian-data-protection-authority-sari-facial-recognition-system-proposed-ministry-interior

Uso de tecnología de reconocimiento facial en vivo en Buenos Aires, Argentina

Entre 2019 y 2022, se implementó tecnología de reconocimiento facial en vivo en Buenos Aires, capital de Argentina, para ayudar a las fuerzas de seguridad a identificar posibles delincuentes buscados en la base de datos nacional de fugitivos del país. El sistema se basó en imágenes en vivo de los sistemas de monitoreo de video ubicados en toda la ciudad, incluidas las tres estaciones de tren principales, y la red de transporte subterráneo, que es utilizada por más de 1,3 millones de pasajeros cada día. Sin embargo, en abril de 2022, se aprobó una orden judicial para suspender temporalmente el uso de la tecnología debido a acusaciones de búsquedas no autorizadas. Y en septiembre de 2022, un tribunal de la ciudad dictaminó que las condiciones actuales en las que operaba el sistema eran inconstitucionales, lo que se espera que extienda aún más la suspensión del sistema de reconocimiento facial. Según la Asociación por los Derechos Civiles (ADC) de Argentina, la tecnología de reconocimiento facial se ha implementado no solo en la capital, sino también en otras regiones, incluidas las provincias de Córdoba, Salta y Mendoza, así como en el partido de Tigre en Buenos Aires. Se ha informado que también hay planes para desplegar la tecnología en la provincia de Santa Fe. Esta información era precisa a principios de 2021.

Uso de la tecnología de reconocimiento facial en Brasil

El uso de la tecnología de reconocimiento facial está bastante extendido en Brasil, con implementaciones identificadas en 30 ciudades a partir de 2019. Esta tecnología se emplea para una variedad de propósitos, incluida la prevención del fraude en la distribución de beneficios sociales. Se ha utilizado para verificar las identidades de los beneficiarios de subsidios de transporte público en numerosas ciudades brasileñas y realizar un seguimiento de los requisitos de asistencia escolar para los programas de transferencia de efectivo en el estado de Pernambuco. Sin embargo, la tecnología de reconocimiento facial también se ha implementado con fines de marketing, como colocar anuncios frente a los pasajeros en el metro de São Paulo utilizando técnicas de detección de emociones altamente controvertidas. Este proyecto finalmente se suspendió después de que un tribunal local declarara que la recopilación de datos sobre los pasajeros del Metro no cumplía con los requisitos mínimos de consentimiento.

Argentina y Brasil son sistemas federales que tienen una compleja coexistencia de leyes municipales, estatales y federales. Esto a menudo conduce a un mosaico de regulaciones con diferentes estándares y salvaguardias que pueden resultar bastante confusas. Esta complejidad ha llevado a desafíos para justificar la legalidad de los despliegues de reconocimiento facial. En Argentina y Brasil, los gobiernos locales han implementado una combinación de legislación municipal y propuestas regulatorias a nivel estatal que a menudo no cumplen con los estándares descritos en sus respectivas constituciones, tratados internacionales de derechos humanos y leyes federales.

Fuente: Chatam House (2022). Regulating facial recognition in Latin America, Policy lessons from police surveillance in Buenos Aires and São Paulo, disponible en: https://www.chathamhouse.org/2022/11/regulating-facial-recognition-latin-america/03-facial-recognition-rollouts-trends-buenos

Clasificación de la protección de datos como un derecho independiente

La clasificación de la protección de datos como un derecho independiente ha sido un punto de discusión en los tribunales internacionales y en el mundo académico. Se deriva del hecho de que la protección de datos, como cuestión regulatoria, surgió en parte de las regulaciones, normas e inquietudes de privacidad, y evolucionó en nuevos conjuntos de obligaciones impuestas a las autoridades públicas y entidades comerciales para proporcionar a las personas el control sobre la información que les concierne, así como los medios para lograr ese control: acceso a esta información, confirmación de su existencia, corrección de datos incorrectos, etc.

Sin embargo, la protección de datos se extiende más allá de las preocupaciones de privacidad. Puede haber problemas importantes de protección de datos cuando las consideraciones de privacidad son irrelevantes o secundarias, como se ilustra a continuación en la sección que trata de los principios de protección de datos.[373] La protección de datos se basa en el derecho a la

373 Ibid.

privacidad, pero también abarca otros derechos de los interesados ante el gobierno y las grandes corporaciones que recopilan, procesan y almacenan datos personales, como el derecho a ser informado, el derecho de acceso a los datos personales, el derecho al olvido, el derecho a la rectificación, el derecho a la portabilidad de los datos, el derecho a oponerse al procesamiento y los derechos relacionados con la toma de decisiones automatizada y la elaboración de perfiles.[374]

Muchos países de todo el mundo reconocen la protección de datos como un derecho fundamental. La protección de datos personales está incorporada como un derecho independiente en varios estatutos, incluida la Carta de los Derechos Fundamentales de la Unión Europea (artículo 8). También fue reconocido recientemente como tal por el Tribunal Supremo de Brasil. Del mismo modo, en un caso reciente (Justice K. S. Puttaswamy (Retd.) vs. Union of India[375]), el Tribunal Supremo de la India afirmó la privacidad como un derecho fundamental.[376]

Derechos de protección de datos relacionados con la toma de decisiones automatizada y la elaboración de perfiles

En muchas jurisdicciones, los interesados tienen derechos relacionados con la toma de decisiones automatizada y la elaboración de perfiles. Esto abarca varias técnicas de elaboración de perfiles, que pueden implicar la evaluación de características personales específicas vinculadas a un individuo que evalúa o pronostica el comportamiento relacionado con el desempeño en el trabajo, la condición financiera, la salud, las preferencias personales, los pasatiempos, la confiabilidad, la conducta o la ubicación. El derecho a estar exento de la toma de decisiones automatizada generalmente se garantiza a los interesados cuando esas decisiones tienen un impacto importante en sus vidas. Sin embargo, estos derechos no se aplican a las decisiones parcialmente automatizadas. Tampoco aseguran necesariamente que, en la práctica, un individuo afectado pueda detectar fácilmente si ha sido tratado de manera desigual con respecto a los demás y, de ser así, si dicho trato diferencial equivalía a discriminación y, por lo tanto, era ilegal. El interesado tiene libertad para renunciar a algunos de sus derechos consintiendo prácticas específicas que de otro modo constituirían una violación de derechos, renunciando así a las protecciones que estos derechos proporcionan.

Por ejemplo, existe un riesgo significativo de que los titulares de derechos individuales renuncien demasiado fácilmente a los derechos de protección de datos en una era en red basada en un modelo comercial de "servicios gratuitos" a cambio del acceso "gratuito" a los servicios digitales y la eficiencia y conveniencia que ofrecen, las personas intercambiarán voluntariamente sus datos personales.[377] Por otro lado, los principios básicos de protección de datos incluyen obligaciones perentorias impuestas a los responsables del tratamiento de datos a las que no pueden renunciar los titulares de derechos individuales, incluidos los principios de licitud

374 Ibid.
375 Status as Fundamental Right (2017). Justice K.S. Puttaswamy (Retd.) vs. Union of India, disponible en: https://privacylibrary.ccgnlud.org/case/justice-ks-puttaswamy-ors-vs-union-of-india-ors
376 UNESCO (2018). Legal Standards on Freedom of Expression, Toolkit for the Judiciary in Africa, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000366340.
377 Comité de Expertos en Intermediarios de Internet (MSI-AUT) (2019). A study of the implications of advanced digital technologies (including AI systems) for the concept of responsibility within a human rights framework, Council of Europe Study, DGI/2019/05, disponible en: https://rm.coe.int/a-study-of-the-implications-of-advanced-digital-technologies-including/168096bdab

del tratamiento, de especificación de la finalidad y de minimización de los datos. Esto ofrece una protección más sistemática y sólida de los valores fundamentales subyacentes y los intereses colectivos que los regímenes de protección de datos en última instancia buscan proteger.[378]

378 Ibid.

Actividad: Los participantes en la capacitación leen los estudios de caso a continuación y debaten cómo se aplican las leyes de protección de datos en sus jurisdicciones, señalando casos de renombre y comparándolos con los casos del RGPD a continuación. ¿Cómo se juzgaría y decidiría un caso similar en su jurisdicción? ¿Qué leyes se aplicarían?

Violaciones de privacidad de Meta en la UE

Después de que se descubriera que la información personal de las personas usuarias de Facebook se publicaba en un foro de hackers en línea, Meta, el propietario de Facebook, fue multado con 265 millones de euros por el regulador irlandés, la Comisión de Protección de Datos, por violar las leyes de protección de datos. La información filtrada incluía los nombres completos, la información de contacto, las fechas de nacimiento y las localidades de las personas usuarias de Facebook en 2018 y 2019.

Meta reconoció que la información se había recopilado utilizando tecnologías destinadas a ayudar a las personas a descubrir amigos mediante números de teléfono. Facebook fue penalizado por "no aplicar la protección de datos por diseño y por defecto" de acuerdo con el RGPD. La multa pudo haberse evitado si esta característica se hubiera diseñado para ser más segura.

Fuente: Satariano A. (2022). Meta Fined $275 Million for Breaking E.U. Data Privacy Law, disponible en: https://www.nytimes.com/2022/11/28/business/meta-fine-eu-privacy.html

Violaciones de privacidad de Google en la UE

El 6 de enero de 2022, la autoridad francesa de protección de datos (CNIL) multó a Google Ireland con 90 millones de euros. La multa se refiere a cómo Google Europe implementa los procesos de consentimiento de cookies de YouTube. La multa de Google Ireland fue una de las dos sanciones emitidas en el mismo caso; la otra se impuso contra Google LLC de California (que opera Google Search).

Google debería haber permitido a las personas usuarias de YouTube rechazar las cookies fácilmente, según la CNIL. YouTube coloca cookies en los dispositivos con fines de marketing para realizar un seguimiento de las actividades en línea. Es fácil aceptar cookies en YouTube, pero es más difícil rechazarlas. La CNIL observó que rechazar cookies requería muchos clics, pero aceptar cookies requería solo uno. Según el RGPD, el consentimiento debe ser "voluntario": si una oferta puede aceptarse con un solo clic, también debería ser posible rechazarla con un solo clic.

La CNIL justificó el castigo comparativamente fuerte citando el gran número de usuarios de YouTube y las enormes ganancias de Google del sitio.

Fuente: Lomas N. (2022). France spanks Google $170M, Facebook $68M over cookie consent dark patterns, disponible en: https://techcrunch.com/2022/01/06/cnil-facebook-google-cookie-consent-eprivacy-breaches/.

Restricciones legítimas al derecho a la privacidad

El Pacto Internacional de Derechos Civiles y Políticos (artículo 2) establece que los Estados partes en el Pacto "respetarán y garantizarán" sin discriminación los derechos enumerados en el Pacto para todas las personas que se encuentren en su territorio y bajo su jurisdicción. Sin embargo, los derechos de privacidad no son absolutos. En muchas jurisdicciones, los organismos encargados de hacer cumplir la ley están exentos de la legislación sobre privacidad de datos[379]. Los gobiernos pueden perturbar legítimamente la privacidad de una persona bajo ciertas circunstancias especificadas por la ley, como emergencias o amenazas a la seguridad nacional. Cualquier limitación de los derechos enumerados en el PIDCP debe estar permitida en virtud de las disposiciones pertinentes del PIDCP. Los gobiernos deben justificar sus acciones de vigilancia y demostrar que cualquier invasión de la privacidad está establecida en leyes y reglamentos que sean claros y precisos, necesarios[380] para lograr objetivos gubernamentales legítimos y proporcionales a la consecución de estos objetivos limitados. Una institución judicial o administrativa independiente, imparcial y competente debe supervisar las acciones de vigilancia de los organismos encargados de hacer cumplir la ley. Además, los funcionarios gubernamentales y otros deben asumir su responsabilidad por mala conducta y errores.[381]

Según la Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos, las actividades de vigilancia estatal deben cumplir la ley. Las excepciones a la vigilancia digital deben limitarse y basarse en los principios de necesidad y proporcionalidad para garantizar una privacidad de datos adecuada en todas las ramas del gobierno.[382] Los siguientes requisitos mínimos deben regir la promulgación de leyes específicas de vigilancia:

- La ley debe ser accesible al público y adecuadamente específica. La ley debe definir con precisión el alcance de la discrecionalidad de vigilancia otorgada a la agencia gubernamental y la forma de vigilancia. La ley también debe describir la naturaleza del delito y la clase de personas que pueden estar sujetas a vigilancia. Las referencias no específicas a "seguridad nacional" o "salud pública" no califican como justificaciones específicas y legítimas, ya que son vagas y amplias. La vigilancia debe basarse en sospechas razonables, y toda decisión de autorizar la vigilancia debe estar suficientemente orientada. La ley debe definir con precisión las competencias de la institución con autoridad para llevar a cabo la vigilancia digital.
- En cuanto a su alcance, el marco legal para la vigilancia también debe incluir las solicitudes de vigilancia del gobierno a las empresas. El marco legal también debe incluir el acceso a la información mantenida extraterritorialmente y el intercambio de información con otros estados.

379 © UNESCO 2022 Guidelines for Judicial Actors on Privacy and Data Protection, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000381298

380 El componente de necesidad de la prueba de restricciones es el más desafiante y controvertido. Involucra varios factores en varias jurisdicciones internacionales. Dos factores clave para determinar la necesidad son (i) la restricción debe servir a una necesidad social urgente, y (ii) las justificaciones de la restricción deben ser suficientes y pertinentes. Véase: Icelandic Human Rights Centre, https://www.humanrights.is/en/human-rights-education-project/comparative-analysis-of-selected-case-law-achpr-iachr-echr-hrc/the-right-to-freedom-of-opinion-and-expression/permissible-limitations. Véase también: Comisión Australiana de Derechos Humanos, Permissible Limitations on Rights, https://humanrights.gov.au/our-work/rights-and-freedoms/permissible-limitations-rights

381 Icelandic Human Rights Centre, https://www.humanrights.is/en/human-rights-education-project/comparative-analysis-of-selected-case-law-achpr-iachr-echr-hrc/the-right-to-freedom-of-opinion-and-expression/permissible-limitations. Véase también: ONU (2018). El derecho a la privacidad en la era digital. Informe del Alto Comisionado de las Naciones Unidas para los Derechos Humanos, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/239/58/PDF/G1823958.pdf?OpenElement

382 Consejo de Derechos Humanos de las Naciones Unidas (2018). El derecho a la privacidad en la era digital. Informe del Alto Comisionado de las Naciones Unidas para los Derechos Humanos, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G18/239/58/PDF/G1823958.pdf?OpenElement

La ley debe establecer explícitamente una estructura para garantizar la rendición de cuentas y la transparencia dentro de las organizaciones gubernamentales que realizan la vigilancia.

- Las facultades de vigilancia solo pueden justificarse si son estrictamente necesarias para alcanzar un objetivo legítimo y si satisfacen el requisito de proporcionalidad. El alcance de la vigilancia debe limitarse a prevenir o investigar los delitos o amenazas más graves. La duración de la vigilancia debe mantenerse al mínimo absoluto requerido para lograr el objetivo especificado. Sobre la base de la estricta necesidad y proporcionalidad, la ley debe contener reglas estrictas para el uso y almacenamiento de los datos recopilados, y definir con precisión las circunstancias en las que los datos recopilados y almacenados deben borrarse. Las mismas reglas de legalidad, estricta necesidad y proporcionalidad deben aplicarse al intercambio de inteligencia.[383]
- Cuando los gobiernos contemplen el hackeo dirigido, deben proceder con extrema cautela, recurriendo a tales medidas solo en circunstancias excepcionales, para la investigación o prevención de los delitos o amenazas más graves, y con la participación del poder judicial. El diseño de las operaciones de piratería debe ser limitado, restringiendo el acceso a objetivos y categorías de información específicos. Los Estados no deben obligar a las entidades privadas a ayudar en las operaciones de piratería, ya que hacerlo comprometería la seguridad de sus propios productos y servicios. El descifrado obligatorio solo se puede permitir caso por caso, con una orden judicial y el mantenimiento de los derechos al debido proceso.[384]

Las medidas de vigilancia, como las solicitudes de datos de comunicaciones de las empresas y el intercambio de inteligencia, deben autorizarse, revisarse y supervisarse por organismos independientes en todas las etapas, incluso cuando se ordenan inicialmente, mientras se llevan a cabo y cuando se dan por terminadas.[385] El organismo independiente que autoriza medidas de vigilancia particulares, preferiblemente una autoridad judicial, debe garantizar que haya pruebas suficientes de una amenaza y que la vigilancia propuesta sea específica, estrictamente necesaria y proporcionada antes de autorizar (o rechazar) las medidas de vigilancia ex ante.

El organismo independiente que autoriza medidas de vigilancia particulares, preferiblemente una autoridad judicial, debe garantizar que haya pruebas claras de una amenaza suficiente y que la vigilancia propuesta sea específica, estrictamente necesaria y proporcionada, y autorizar (o rechazar) ex ante las medidas de vigilancia.[386]

383 Consejo de Derechos Humanos de las Naciones Unidas (2013). Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión, disponible en: https://www.ohchr.org/Documents/HRBodies/HRCouncil/RegularSession/Session23/A.HRC.23.40_EN.pdf

384 Consejo de Derechos Humanos de las Naciones Unidas (2015). Informe del Relator Especial sobre la promoción y protección del derecho a la libertad de opinión y de expresión, disponible en: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G15/095/85/PDF/G1509585.pdf?OpenElement

385 Pacto Internacional de Derechos Civiles y Políticos. (2015). Concluding observations on the fifth periodic report of France, disponible en: https://tbinternet.ohchr.org/_layouts/15/TreatyBodyExternal/Download.aspx?symbolno=CCPR%2FC%2FFRA%2FCO%2F5&Lang=en

386 Agencia Europea de Derechos Fundamentales (2017). Vigilancia por parte de los servicios de inteligencia: garantías y recursos de los derechos fundamentales en la UE. Volumen II: Field Perspectives and Legal Update, disponible en: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2017-surveillance-intelligence-services-vol-2_en.pdf

Los marcos de supervisión incluyen agencias administrativas, judiciales y/o parlamentarias. Los órganos de supervisión deben ser independientes de las autoridades de vigilancia y estar dotados de la experiencia, las habilidades y los recursos necesarios. Institucionalmente, las reglas deben diferenciar y separar las funciones de autorización y supervisión. Además de las evaluaciones periódicas de las capacidades de vigilancia y los avances tecnológicos, los organismos de supervisión independientes deben investigar y monitorear las actividades de quienes realizan la vigilancia y acceden a sus productos.[387] Se debe exigir a las agencias que realizan la vigilancia que proporcionen toda la información necesaria para una supervisión efectiva previa solicitud, que presenten informes periódicos a los organismos de supervisión y mantengan registros de todas las medidas de vigilancia. Además, los procesos de supervisión deben ser abiertos y estar sujetos al escrutinio público apropiado, y las decisiones de los órganos de supervisión deben estar sujetas a apelación o revisión independiente.[388]

Principio de transparencia: el debate y el escrutinio abiertos son cruciales para comprender los beneficios y las limitaciones de las técnicas de vigilancia, por lo tanto, las autoridades estatales y los organismos de supervisión también deben participar en la información pública sobre las leyes, políticas y prácticas existentes en materia de vigilancia e interceptación de comunicaciones, así como otras formas de procesamiento de datos personales.[389] El organismo de vigilancia debe explicar la limitación del derecho a la privacidad a aquellos que fueron objeto de vigilancia. Además, las personas sometidas a vigilancia deben tener derecho a cambiar y eliminar información personal innecesaria si ya no es requerida para investigaciones en curso o futuras.[390]

En principio, para ser legales, las restricciones al derecho a la privacidad a través del marco nacional de derechos humanos, la protección de datos, la ciberseguridad, el delito cibernético y la vigilancia digital o las leyes y políticas de TIC deben cumplir con ciertos estándares mínimos del derecho internacional de los derechos humanos. Estos estándares se pueden encontrar en la Resolución de la Asamblea General de la ONU sobre el Derecho a la Privacidad en la Era Digital de 2014[391], el Informe de 2014 del Relator Especial sobre la Promoción y Protección de los Derechos Humanos y las Libertades Fundamentales en la[392] Lucha contra el Terrorismo y el Informe de la Oficina del Alto Comisionado de las Naciones Unidas para el Derecho Humano a la Privacidad en la Era Digital[393]. De acuerdo con estos estándares[394], para ser legales, las restricciones al derecho a la privacidad hechas por los gobiernos deben ser:

387 Véase Tribunal Europeo de Derechos Humanos, Kennedy vs. Reino Unido, solicitud N.º 26839/05, sentencia del 18 de mayo de 2010.
388 https://www.cipil.law.cam.ac.uk/projects/human-rights-big-data-and-technology-hrbdt-project
389 Consejo de Derechos Humanos de las Naciones Unidas (2009). Informe del Relator Especial sobre la promoción y protección de los derechos humanos y las libertades fundamentales en la lucha contra el terrorismo, disponible en: https://daccess-ods.un.org/tmp/9699321.3891983.html
390 Consejo de Derechos Humanos de las Naciones Unidas (2017). Informe del Relator Especial sobre el derecho a la privacidad, disponible en: https://daccess-ods.un.org/tmp/2525206.50625229.html
391 Consejo de Derechos Humanos de las Naciones Unidas (2014). The right to privacy in the digital age: report of the Office of the United Nations High Commissioner for Human Rights, disponible en: https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F %2Fwww.ohchr.org%2Fsites%2Fdefault%2Ffiles%2FDocuments%2FIssues%2FDigitalAge%2FA-HRC-27-37_en.doc&wdOrigin=BROWSELINK
392 Véase: https://www.ohchr.org/en/special-procedures/sr-terrorism
393 Consejo de Derechos Humanos de las Naciones Unidas (2021). The right to privacy in the digital age, Report of the United Nations High Commissioner for Human Rights, disponible en: https://www.ohchr.org/en/HRBodies/HRC/RegularSessions/Session48/Documents/A_HRC_48_31_AdvanceEditedVersion.docx
394 Cabe señalar que estos estándares no son aceptados universalmente por todos los gobiernos. Muchos gobiernos interpretan las disposiciones del PIDCP de manera diferente. Por ejemplo, Estados Unidos ha señalado históricamente (consulte la página 235, disponible en: https://2017-2021.state.gov/wp-content/uploads/2019/10/2018-Digest-Final-Draft.pdf#page =235) que el Artículo 19 del Pacto Internacional de Derechos Civiles y Políticos no impone un estándar de legalidad, necesidad y proporcionalidad, solo que la vigilancia no puede ser ilegal o arbitraria.

- ➔ **Impuestas solo para proteger propósitos legítimos:** con respecto al derecho a la privacidad, la vigilancia digital solo debe autorizarse en pos de los objetivos nacionales más vitales. La restricción debe ser esencial para lograr un objetivo legítimo, proporcional al objetivo y la opción menos invasiva disponible. Además, debe demostrarse que la restricción impuesta al derecho (como una invasión de la privacidad para salvaguardar la seguridad nacional o el derecho a la vida de otros) puede lograr razonablemente su objetivo previsto. La carga recae en las autoridades que intentan restringir el derecho a demostrar que la restricción sirve a un objetivo legítimo.
- ➔ **Legítimas:** los límites al derecho a la privacidad deben establecerse de manera clara e inequívoca en la ley y deben revisarse con frecuencia para garantizar que las protecciones y salvaguardias de la privacidad se mantengan al ritmo de los rápidos desarrollos de la tecnología digital.[395] Según el Informe de la Oficina del Alto Comisionado de las Naciones Unidas para los Derechos Humanos, "El derecho a la privacidad en la era digital": "la interferencia que está permitida por la legislación nacional puede, sin embargo, ser "ilegal" si esa legislación nacional está en conflicto con las disposiciones del Pacto Internacional de Derechos Civiles y Políticos".[396]
- ➔ **Cumplir con el principio de no discriminación en su diseño y aplicación:** los límites al derecho a la privacidad no deben discriminar a ningún grupo vulnerable.
- ➔ **Necesaria y proporcionada:** la vigilancia digital es un acto muy intrusivo que vulnera el derecho a la intimidad. Es necesaria la aprobación previa de una autoridad judicial competente para una vigilancia digital proporcionada. Esto también significa que se utilizarán los métodos de vigilancia menos intrusivos.[397]

Los gobiernos limitan el derecho a la privacidad por las siguientes razones:

- Seguridad nacional
- Seguridad pública
- Bienestar económico nacional
- Protección de los derechos y libertades de los demás
- Prevención del desorden o el delitoe
- Protección de la salud o la moral[398]

395 MISA Zimbabwe, Konrad Adenauer Stiftung (2021). Cybersecurity and Cybercrime Laws in the SADC Region: Implications on Human Rights, disponible en: https://fdocuments.net/document/cybersecurity-and-cybercrime-laws-in-the-sadc-region.html?Página 3

396 Consejo de Derechos Humanos de las Naciones Unidas (2014). The right to privacy in the digital age: report of the Office of the United Nations High Commissioner for Human Rights, disponible en: https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.ohchr.org%2Fsites%2Fdefault%2Ffiles%2FDocuments%2FIssues%2FDigitalAge%2FA-HRC-27-37_en.doc&wdOrigin=BROWSELINK%20

397 International Commission of Jurists, Regulation of Communications Surveillance and Access to Internet in Selected African States, disponible en: https://www.kas.de/documents/275350/0/Report-on-Regulation-of-Communications-Surveillance-and-Access-to-Internet-in-Selected-African-States.pdf/66dbd47d-4d7d-2779-a595-a34e9f93cfbb?t=1639140695434

398 Véase: https://africaninternetrights.org/en/node/2558#:~:text=This%20advocacy%20toolkit%20provides%20an%20overview%20of%20the,the%20formulation%20and%20implementation%20of%20data%20protection%20frameworks.

En profundidad:

Enfoque basado en los derechos humanos (HRBA) para evaluar el impacto de la regulación sobre el derecho a la privacidad en el entorno digital

Los países que promulgan leyes sobre ciberdelincuencia, ciberseguridad y protección de datos deben seguir al HRBA en la redacción de la regulación digital. Un HRBA se basa en los principios derivados de los tratados internacionales y regionales y coloca los derechos humanos en el centro de toda formulación de políticas y redacción legislativa. Los elementos fundamentales de este enfoque son la participación, la responsabilidad y la transparencia, la no discriminación y la igualdad, el empoderamiento de los titulares de derechos y la legalidad. La regulación de la vigilancia digital debe ser inequívoca con respecto a qué agencias pueden llevar a cabo la vigilancia, quién puede juzgar las solicitudes de vigilancia, qué pruebas legales debe aplicar un tribunal a las solicitudes y qué sanciones legales se aplican a la vigilancia no autorizada.[399] Los abogados y grupos de defensa y las OSC que trabajan en el área de la privacidad digital deben utilizar la HBRA como una herramienta para evaluar si las restricciones impuestas al derecho a la privacidad por el gobierno son legítimas, legales, compatibles con el principio de no discriminación en su diseño y aplicación, y necesarias y proporcionadas.[400]

Una HBRA debe implicar una evaluación de la regulación digital nacional en relación con los **Principios Internacionales sobre la Aplicación de los Derechos Humanos a la Vigilancia de las Comunicaciones**[401]. La siguiente figura describe las áreas clave en las que se centran estos Principios:

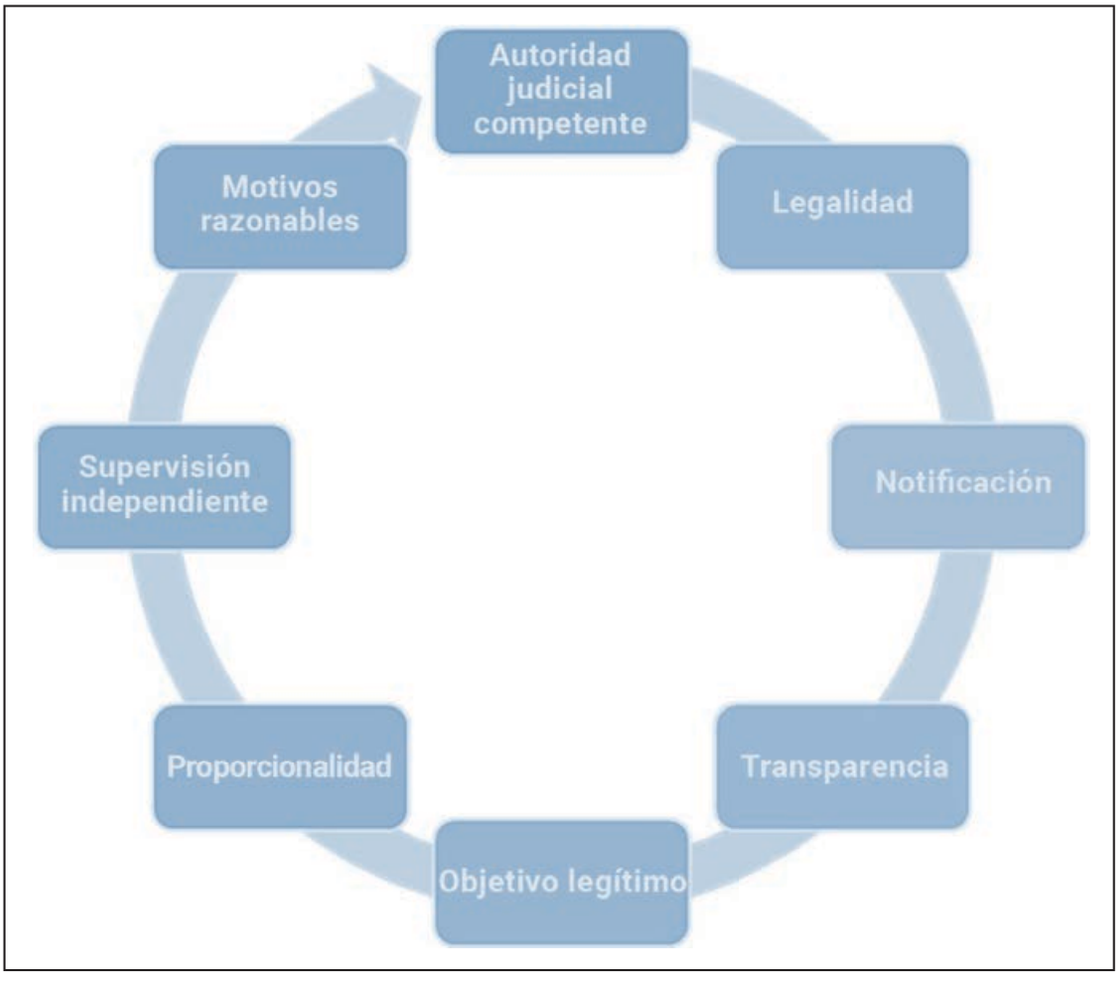
- **Autorización previa de vigilancia por parte de una autoridad judicial competente:** ¿existe un juez con experiencia en tecnología digital y derechos humanos que pueda evaluar y autorizar las solicitudes de vigilancia de las agencias gubernamentales investigadoras?
- **Objetivo legítimo:** ¿la ley establece ciertos fines lícitos de vigilancia, como la prevención del terrorismo o delitos graves con una pena judicial de 10 o más años de cárcel?
- **Motivos razonables:** ¿los jueces están facultados para determinar si existe un alto nivel de amenaza para un objetivo legítimo y una alta probabilidad de que la vigilancia genere pruebas que eliminen la amenaza?
- **Legalidad:** ¿la vigilancia se lleva a cabo exclusivamente dentro de las limitaciones y por los organismos especificados por la ley? ¿La ley declara ilegal cualquier otra vigilancia y estipula sanciones?
- **Necesidad:** ¿están los jueces autorizados a determinar si se requiere monitoreo para asegurar la evidencia y que no existe un método menos intrusivo para lograr el propósito legítimo?
- **Proporcionalidad:** ¿los jueces están facultados para determinar si la vigilancia propuesta tiene un alcance limitado y la duración es proporcional a las pruebas necesarias para eliminar la amenaza?

399 Véase: https://unsdg.un.org/2030-agenda/universal-values/human-rights-based-approach

400 Roberts T., Mohamed A., Farahat, M., Oloyede R., Mutung'u G. (2021). Surveillance Law in Africa: a Review of Six Countries, Institute of Development Studies: Brighton, disponible en: https://opendocs.ids.ac.uk/opendocs/bitstream/handle/20.500.12413/16893/Roberts_Surveillance_Law_in_Africa.pdf

401 Más de 600 grupos, incluidos Privacy International, Open Rights Group, Electronic Frontier Foundation y la Asociación para el Progreso de las Comunicaciones, coordinaron la redacción de los Principios Internacionales, disponible en: https://www.eff.org/files/necessaryandproportionatefinal.pdf

- **Notificación al sujeto:** ¿la ley requiere que el sujeto de vigilancia sea informado de la vigilancia lo antes posible para brindar una oportunidad de apelación legal y debido proceso?
- **Informes de transparencia**: ¿los informes anuales sobre apertura hacen pública la cantidad de solicitudes, justificaciones y autorizaciones de vigilancia?
- **Supervisión independiente**: ¿las prácticas de vigilancia tienen algún mecanismo de monitoreo público para garantizar su responsabilidad y transparencia?[402]



Fuente: Adapted from Roberts T., Mohamed A., Farahat, M., Oloyede R., Mutung'u G. (2021). Surveillance Law in Africa: a Review of Six Countries, Brighton: Institute of Development Studies, available at: DOI: 10.19088/IDS.2021.059

402 Roberts T., Mohamed A., Farahat, M., Oloyede R., Mutung'u G. (2021). Surveillance Law in Africa: a Review of Six Countries, Institute of Development Studies: Brighton, disponible en: https://opendocs.ids.ac.uk/opendocs/bitstream/handle/20.500.12413/16893/Roberts_Surveillance_Law_in_Africa.pdf

3. Enfoques para la gobernanza de la IA

A medida que la IA se integra rápidamente en todos los sectores, es importante que los operadores judiciales consideren los beneficios y riesgos únicos asociados con los diferentes sistemas de IA. Los asistentes virtuales, los vehículos autónomos y las recomendaciones en vídeo para niños presentan diferentes niveles de beneficios y riesgos. Por lo tanto, la formulación de políticas y la gobernanza deben abordarse de manera diferente para cada sistema de IA específico en función de los riesgos involucrados, su gravedad y su impacto en los derechos humanos. La Tabla 7 a continuación ofrece una descripción general de los principios rectores que rigen la IA.

Tabla 7. Seleccionar los principios rectores que rigen la IA

Principios	Cuestiones importantes en la implementación de los principios
Cuanto mayor sea el riesgo para los derechos humanos, más estrictas deben ser las normas legales para el uso de la tecnología de IA.	Los sectores en los que hay mucho en juego para la invasión de los derechos fundamentales individuales, como la seguridad nacional, la justicia penal, la aplicación de la ley, la salud y la protección social, deben tener prioridad. Un enfoque proporcional al riesgo para la regulación de la IA requerirá la prohibición de tecnologías, aplicaciones y casos de uso específicos de IA que produzcan impactos potenciales o reales que violen los derechos humanos internacionales, incluidos aquellos que no cumplan con los requisitos de necesidad y proporcionalidad.[403]
No se deben permitir aplicaciones de IA que discriminen.	Debería prohibirse la calificación social de las personas por parte de los gobiernos[404] o el uso de sistemas de IA que clasifiquen a las personas en grupos basados en factores discriminatorios prohibidos. Los gobiernos deberán controlar el uso y la adquisición de tecnologías de IA cuyo despliegue en el poder judicial plantea peligros para los derechos humanos. Cuando es probable que ocurran violaciones de los derechos humanos, se debe exigir el requisito de monitoreo humano (humano en el circuito). Los gobiernos deben posponer el despliegue de tecnologías potencialmente de alto riesgo, como el reconocimiento facial remoto en tiempo real, hasta que se pueda garantizar que su implementación no violará los derechos humanos.[405]
Si se utiliza un sistema de IA para interactuar con humanos en el contexto de los servicios públicos, en particular la justicia, el bienestar y la atención médica, se debe informar e informar al usuario de la opción de consultar a un profesional previa solicitud y sin demora.	Aquellos que han recibido una decisión tomada por una autoridad pública que se basa única o sustancialmente en la salida de un sistema de IA deben ser alertados y recibir la información antes mencionada lo más pronto posible.[406] Esto puede consistir en la divulgación pública de información sobre el sistema en cuestión, sus procesos, los efectos directos e indirectos sobre los derechos humanos y las medidas adoptadas para identificar y mitigar las consecuencias adversas del sistema para los derechos humanos, o en una auditoría imparcial, exhaustiva y eficaz. En todos los casos, la información proporcionada debe permitir una evaluación significativa del sistema de IA. Ningún sistema de IA debe ser tan complicado que la evaluación e inspección humanas sean imposibles. Los sistemas de ADM que no pueden cumplir con los estándares adecuados de transparencia y responsabilidad no deben utilizarse en la prestación de servicios públicos.[407]

403 La propuesta de Ley de IA de la Unión Europea adopta un enfoque basado en el riesgo.
404 CAHAI (2020). The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law, párr. 75, disponible en: https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da; véase también: UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence, disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000380455
405 Parlamento Europeo (2019). A governance framework for algorithmic accountability and transparency, disponible en:https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2019)624262
406 CAHAI (2020). The Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law, disponible en: https://rm.coe.int/cahai-2020-06-fin-c-muller-the-impact-of-ai-on-human-rights-democracy-/16809ed6da
407 Ibid.

El proyecto de ley de IA de la UE como ejemplo de regulación basada en el riesgo de la IA

En 2019, tras la publicación de las Directrices éticas para una IA fiable[408], la Comisión Europea inició un enfoque múltiple para regular la IA y abordar los riesgos relacionados con la IA. Además del proyecto de Ley de IA, las normas de responsabilidad civil nuevas y modificadas[409] actúan junto con otras políticas actuales y previstas relacionadas con los datos, como el RGPD[410], la Ley de Servicios Digitales[411], la Ley de Datos[412] propuesta y la Ley de Resiliencia Cibernética propuesta[413].

El proyecto de Ley de IA de la UE establece estándares horizontales para el desarrollo, la comercialización y el uso de productos, servicios y sistemas impulsados por IA dentro de la UE. Proporciona directrices fundamentales basadas en el riesgo de IA aplicables a todas las industrias e incluye un "marco de seguridad del producto" con cuatro categorías de riesgo, que especifica las reglas de entrada al mercado y la certificación para los sistemas de IA de alto riesgo a través de un proceso obligatorio de marcado CE. Este régimen de cumplimiento también abarca los conjuntos de datos utilizados para la capacitación, las pruebas y la validación del aprendizaje automático para garantizar resultados justos.

El proyecto de ley de IA de la UE emplea una estrategia basada en el riesgo con múltiples mecanismos de aplicación. Las aplicaciones de IA de bajo riesgo estarían sujetas a un marco regulatorio más indulgente, mientras que aquellas con riesgos inaceptables estarían prohibidas. A medida que aumenta el riesgo, se aplican regulaciones más estrictas. Estas varían desde requisitos de certificación externa más ligeros a lo largo del ciclo de vida de la aplicación hasta evaluaciones de impacto de ley blanda autorreguladora no vinculantes combinadas con códigos de conducta.

El marco regulatorio define cuatro niveles de riesgo en IA:

(i) Riesgo inaceptable. Se prohibirán los sistemas de IA perjudiciales para los derechos, la seguridad y los medios de vida de las personas, incluidos los sistemas de puntuación social utilizados por los gobiernos y los juguetes activados por voz que promueven comportamientos de riesgo.[414]

(ii) Riesgo alto. La propuesta inicial (2021) incluía (i) infraestructura fundamental (por ejemplo, transporte), que podría poner en riesgo la vida y la salud de los ciudadanos; formación educativa o profesional que puede determinar el acceso a la educación y el curso profesional de la vida de alguien (por ejemplo, calificación de exámenes);

408 Comisión Europea (2019). Ethics guidelines for trustworthy AI, disponible en: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
409 Comisión Europea (2022). Nuevas normas de responsabilidad sobre productos e IA para proteger a los consumidores y fomentar la innovación, disponible en: https://ec.europa.eu/commission/presscorner/detail/en/ip_22_5807
410 Comisión Europea (2021). Data Protection, disponible en: https://commission.europa.eu/law/law-topic/data-protection_en
411 Comisión Europea (2022). Ley de Servicios Digitales, disponible en: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package
412 Comisión Europea (2023). Data Act: Commission welcomes political agreement on rules for a fair and innovative data economy, disponible en:https://ec.europa.eu/commission/presscorner/detail/en/ip_23_3491
413 Comisión Europea (2022). Cyber Resilience Act , disponible en: https://digital-strategy.ec.europa.eu/en/library/cyber-resilience-act
414 Véase: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

(iii) componentes de seguridad de los productos (por ejemplo, aplicaciones de IA en cirugía asistida por robot); (iv) empleo, gestión de trabajadores y acceso al trabajo por cuenta propia (por ejemplo, reanudar los servicios de clasificación con fines de contratación); (v) servicios públicos y privados esenciales (por ejemplo, calificación crediticia que niega a los ciudadanos la oportunidad de obtener un préstamo); (vi) actividades de aplicación de la ley que interfieren con los derechos humanos (por ejemplo, evaluación de la admisibilidad de pruebas); (vii) gestión de la migración, el asilo y el control fronterizo (por ejemplo, verificación de la autenticidad de los documentos de viaje); (viii) administración de justicia y procesos democráticos (por ejemplo, aplicación de la ley a un conjunto concreto de hechos).

La propuesta de diciembre de 2022 eliminó la detección de falsedades por parte de las fuerzas del orden, el análisis de delitos y la verificación de la autenticidad de los documentos de viaje de la lista de sistemas de IA de alto riesgo. Los últimos cambios aclaran que el alcance del proyecto de ley no abarca la IA para fines de seguridad nacional, defensa y militares.

Todas las tecnologías de identificación biométrica remota están sujetas a estrictas regulaciones y se consideran de alto riesgo. En general, está prohibido emplear la identificación biométrica remota para la aplicación de la ley en áreas abiertas al público. Solo se pueden permitir unas pocas situaciones como excepciones, como cuando es imperativo encontrar a un niño desaparecido, detener una amenaza terrorista específica e inminente, o encontrar, identificar o enjuiciar a un perpetrador o sospechoso de un delito grave. Dicho uso está sujeto a limitaciones de tiempo, ubicación y búsqueda en la base de datos, así como a la aprobación de un órgano judicial u otro órgano imparcial.[415]

(ii) Riesgo limitado. Los sistemas de IA con riesgo limitado deben cumplir con los requisitos de divulgación específicos. Las personas usuarias deben ser conscientes de que están interactuando con una máquina cuando utilizan sistemas de IA como los chatbots para que puedan decidir por sí mismos si avanzan o retroceden.[416]

(iii) Riesgo mínimo o sin riesgo. En esto se incluyen aplicaciones como filtros de spam o videojuegos con IA.

Las personas usuarias aseguran el control y la supervisión humana una vez que se pone en el mercado un sistema de IA, mientras que los proveedores cuentan con una estructura de supervisión posterior a la comercialización. Las autoridades están a cargo de la supervisión del mercado. Tanto los proveedores como las personas usuarias informarán de eventos graves y fallos de funcionamiento.[417]

415 Ibid.
416 Ibid.
417 Ibid.

Enfoques basados en los derechos humanos para la gobernanza de la IA

Un enfoque basado en los derechos humanos es esencial para construir sistemas de IA confiables en la prestación de servicios públicos. Para garantizar un enfoque basado en los derechos en las operaciones del sector público, los gobiernos de los países en desarrollo deben tener un marco analítico fácilmente accesible para ayudarlos a identificar cuándo los componentes de IA podrían afectar los derechos humanos y cómo la responsabilidad algorítmica podría mitigar esos riesgos. Cuando los sistemas de IA amenazan los derechos fundamentales, los países deben proteger y promover esos derechos y garantizar que los actores del sector privado lleven a cabo la diligencia debida y las evaluaciones de impacto en los derechos humanos (HRIA) de acuerdo con su responsabilidad. El resultado de las HRIA debe conducir a diferentes salvaguardias asignadas a los riesgos e impactos específicos establecidos en el proceso ().[418]

Los gobiernos de todo el mundo, como los Estados Unidos (Anteproyecto para una Carta de derechos de la IA),[419] han intentado abordar los problemas de responsabilidad y transparencia de la IA a través de una perspectiva de derechos humanos. Un marco valioso para realizar evaluaciones de impacto algorítmicas basadas en el enfoque de derechos humanos es la herramienta de evaluación de impacto de derechos fundamentales y algoritmos (FRAIA) desarrollada por el Ministerio del Interior y Relaciones del Reino de los Países Bajos.[420]

418 Agencia de los Derechos Fundamentales de la Unión Europea (2022). Sesgo en Algoritmos – Inteligencia Artificial y Discriminación, disponible en: https://fra.europa.eu/sites/default/files/fra_uploads/fra-2022-bias-in-algorithms_en.pdf

419 Casa Blanca (2022). Blueprint for an AI Bill of Rights, disponible en: https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf

420 Ministerio del Interior y Relaciones del Reino de los Países Bajos (2022). Evaluación de impacto Derechos fundamentales y algoritmos, disponible en: https://www.government.nl/documents/reports/2022/03/31/impact-assessment-fundamental-rights-and-algorithms.

Evaluaciones de impacto sobre los derechos humanos (HRIA)

Las HRIA pueden ayudar a identificar riesgos que los operadores judiciales podrían no prever en el desarrollo y despliegue de la IA. Para lograr esto, las HRIA priorizan las implicaciones de los derechos humanos sobre la optimización de la tecnología o sus resultados. Las HRIA o procesos comparables podrían garantizar el respeto de los derechos humanos por diseño a lo largo del ciclo de vida de la tecnología.

Las HRIA evalúan la tecnología en función de una amplia variedad de posibles impactos en los derechos humanos. Cuando se utiliza ADM en los entornos judiciales, las partes interesadas deben llevar a cabo HRIA transparentes, imparciales e inclusivas, que consisten en un examen de los productos, servicios y sistemas de intermediarios que rodean el desarrollo y despliegue de IA y sus efectos en los derechos humanos. Estas HRIA deben incorporar aportes de las comunidades afectadas y las organizaciones de partes interesadas, incluida la sociedad civil y los grupos marginados. Los resultados de las HRIA deben hacerse públicos y deben ser de libre acceso y comprensibles.[421]

Las HRIA para IA deben investigar el funcionamiento interno de los algoritmos, es decir, deben analizar sus componentes técnicos. Las HRIA para algoritmos también deben realizarse a lo largo de todo el ciclo de vida de un sistema de IA, comenzando con las primeras etapas de su concepción y continuando a través de fases importantes de su desarrollo e implementación. No deben ser solo esfuerzos ex ante o ex post. La Evaluación de Impacto de Derechos Fundamentales y Algoritmos (FRAIA), desarrollada por el gobierno holandés, y la Evaluación de Impacto Humano, Ético y Social en IA, desarrollada por Alessandro Manterelo en la Universidad de Turín, son HRIA recientes que cumplen con estos requisitos. Ambas HRIA ofrecen recomendaciones para ayudar a los desarrolladores e implementadores de IA a identificar el impacto de los sistemas de IA en una amplia gama de derechos básicos. Además, proporcionan varias instancias de posibles estrategias de mitigación para prevenir efectos adversos. Todo esto minimiza la probabilidad de violaciones injustificables de los derechos humanos. La FRAIA considera el impacto de los sistemas de IA en más de un centenar de derechos y subderechos fundamentales, por ejemplo, la libertad de expresión se subdivide en numerosos subderechos, como la libertad de prensa, la libertad académica y la denuncia de irregularidades, y propone una lista exhaustiva de medidas preventivas y de mitigación para limitar las infracciones de estos derechos.

A continuación se muestra una instantánea del proceso FRAIA:

Esta Evaluación de Impacto de Derechos Fundamentales y Algoritmos ("FRAIA", por sus siglas en inglés) es una herramienta de debate y de toma de decisiones para organizaciones gubernamentales. La herramienta facilita un diálogo interdisciplinario por parte de los responsables del desarrollo y/o uso de un sistema algorítmico. El cliente encargado es el principal responsable de la implementación (delegada) de la FRAIA.

La FRAIA comprende una gran cantidad de preguntas sobre los temas que es necesario debatir y a las cuales se debe formular una respuesta en cualquier instancia en la que una organización gubernamental considere desarrollar, delegar el desarrollo, comprar, ajustar y/o utilizar un algoritmo (en adelante, en aras de la brevedad: usar o el uso de un algoritmo). Incluso cuando ya se está utilizando un algoritmo, la FRAIA puede servir como herramienta de reflexión. El debate sobre las diversas cuestiones debe llevarlo a cabo un equipo multidisciplinario formado por personas con una amplia gama de especializaciones y formaciones. Por pregunta, la FRAIA indica quiénes deben participar en el debate. Esta herramienta presta atención a todos los roles en un equipo multidisciplinario, que se incluyen en el siguiente diagrama. Sin embargo, la lista no es exhaustiva. Asimismo, los nombres de roles o funciones pueden diferir de una organización a otra.

Rol	FRAIA Parte 1	FRAIA Parte 2	FRAIA Parte 3	FRAIA Parte 4
Grupo de interés	•			
Administración	•			
Panel ciudadano	•			
CIO o CISO	•	•		
Especialista en comunicación		•	•	
Científico de datos		•	•	
Controlador de datos		•		
Experto de dominio (empleado que posee conocimiento del dominio relacionado con el campo de aplicación del algoritmo)	•	•	•	•
Oficial de protección de datos		•		
Personal de RR. HH			•	
Asesor legal	•	•	•	•
Desarrollador de algoritmos		•		
Cliente encargado	•	•	•	
Otros miembros del equipo de proyectos	•			
Líder de proyectos	•	•	•	•
Consultor de ética estratégica		•	•	

Fuente: OCDE, AI in Society, disponible en: https://www.oecd-ilibrary.org/sites/969ff07f-en/index.html?itemId=/content/component/969ff07f-en; Gaumond E., Régis C. (2023). Assessing Impacts of AI on Human Rights: It's Not Solely About Privacy and Nondiscrimination, disponible en: https://www.lawfareblog.com/assessing-impacts-ai-human-rights-its-not-solely-about-privacy-and-nondiscrimination.

421 OSCE (2022). Spotlight on Artificial Intelligence and Freedom of Expression: A Policy Manual, disponible en: https://www.osce.org/representative-on-freedom-of-media/510332

El Marco de Garantía de los Derechos Humanos, la Democracia y el Estado de derecho (HUDERAF) para los sistemas de IA

El HUDERAF, propuesto por el instituto Alan Turing (que ha estado asesorando al CAHAI, el Comité Ad hoc sobre Inteligencia Artificial del Consejo de Europa) tiene como objetivo presentar un método coherente e integrado para evaluar los posibles efectos negativos sobre los derechos humanos, la democracia y el Estado de derecho causados por el uso de sistemas de IA, así como para garantizar que los riesgos identificados que plantea la IA a los operadores judiciales se mitiguen y gestionen adecuadamente. El marco se compone específicamente de varios procedimientos y herramientas bien articulados pero conectados lógicamente. Combina enfoques transparentes de gestión de riesgos, mitigación de impactos y garantía de innovación con evaluaciones de riesgos basadas en el contexto y la participación adecuada de las partes interesadas. Los operadores judiciales podrían utilizar el marco de HUDERAF para evaluar los posibles impactos negativos de la IA en los derechos humanos.

El HUDERAF abarca cuatro componentes:

(1) El Análisis Preliminar de Riesgos Basado en el Contexto (PCRA) ofrece una primera indicación de los riesgos basados en el contexto que un sistema de IA puede representar para los derechos humanos, la democracia y el Estado de derecho. El objetivo principal de la PCRA es ayudar a los equipos de proyectos de IA a desarrollar una estrategia razonable para la gestión de riesgos y los procedimientos de aseguramiento, así como el grado de participación de las partes interesadas requerido a lo largo del ciclo de vida del proyecto.

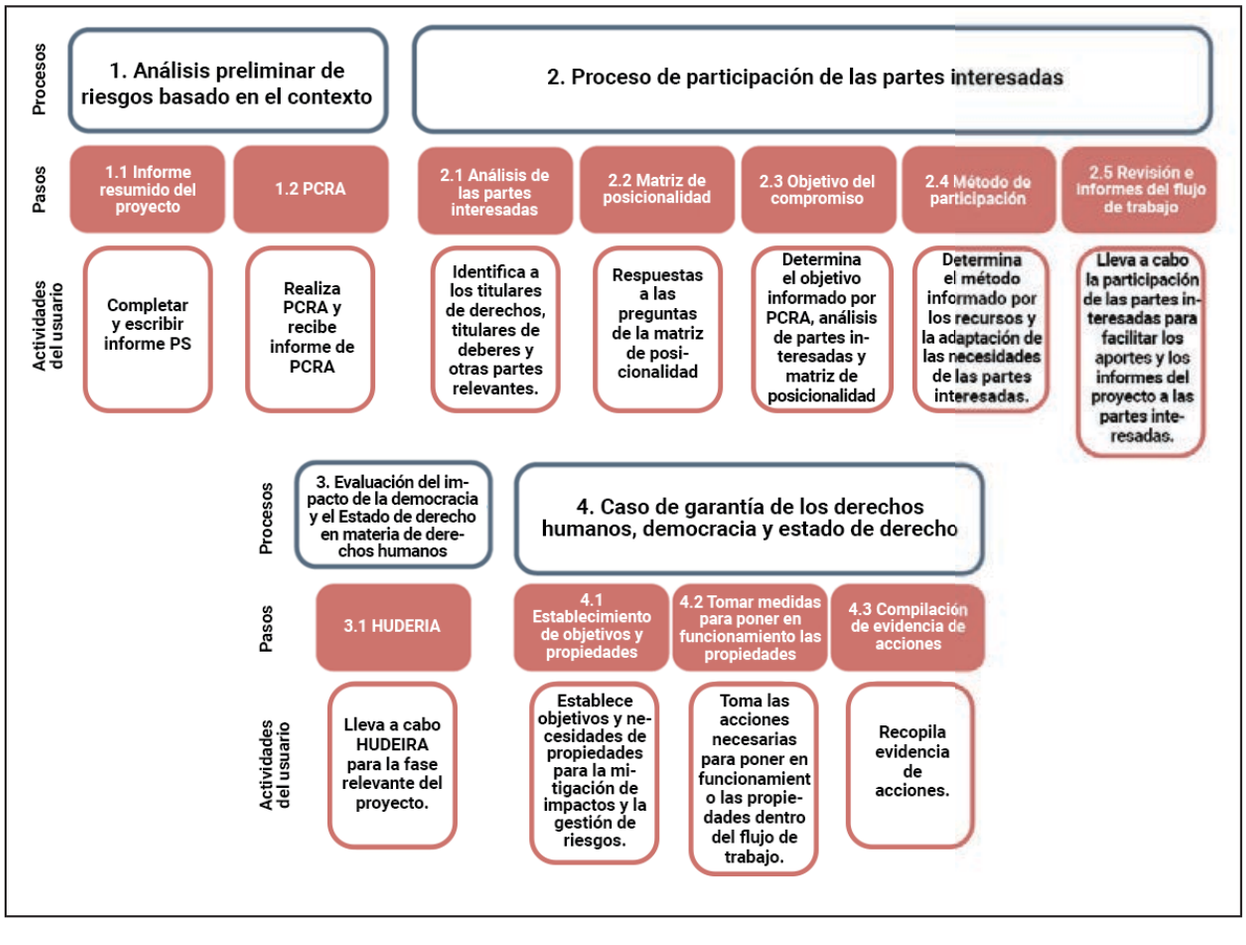
(2) El Proceso de participación de las partes interesadas (SEP) apoya la participación y los aportes apropiados de las partes interesadas a lo largo del proceso del proyecto al ayudar a los equipos del proyecto a identificar la relevancia de las partes interesadas. A través de la participación, la revisita y la revisión de las partes interesadas, este método protege la igualdad y la precisión contextual de los procesos de gobernanza de HUDERAF.

(3) La Evaluación de Impacto de los Derechos Humanos, la Democracia y el Estado de Derecho (HUDERIA) brinda a los equipos del proyecto y a las partes interesadas involucradas la oportunidad de trabajar unidos para crear evaluaciones en profundidad de los efectos posibles y reales que el diseño, desarrollo y uso de un sistema de IA podría tener en los derechos humanos, la democracia y el Estado de derecho. A través de la integración de las perspectivas de las partes interesadas, este proceso contextualiza y valida los daños potenciales previamente identificados, permite el descubrimiento de daños adicionales, permite la evaluación colaborativa de la gravedad de los impactos adversos potenciales identificados, facilita el codiseño de un plan de mitigación de impactos, establece el acceso a la reparación y establece protocolos para el monitoreo y reevaluación de impactos.

4) El Caso de Garantía de Derechos Humanos, Democracia y Estado de derecho (HUDERAC) permite a los equipos de proyectos de IA construir una justificación estructurada que brinde a las partes interesadas una garantía demostrable de que las afirmaciones sobre el logro de los objetivos

establecidos en HUDERIA y otros procesos de gobernanza de HUDERAF están justificadas a la luz de la evidencia disponible. Crear un caso de aseguramiento facilita la reflexión y el debate internos, fomentando la adopción de mejores prácticas e incorporándolas en los ciclos de vida de diseño, desarrollo e implementación. Además, ofrece un medio claro para notificar a las partes interesadas afectadas sobre los pasos tomados a lo largo del flujo de trabajo del proyecto para reducir los riesgos y garantizar la identificación de los objetivos normativos pertinentes. Un caso de aseguramiento cuidadosamente construido proporciona un marco transparente y fácil de entender para gestionar los riesgos y mitigar sus efectos, apoyando los niveles adecuados de aceptación social, responsabilidad y apertura.[422]

Figura 15: Marco para la garantía de los derechos humanos, la democracia y el estado de derecho (HUDERAF)



Fuente: Leslie D., Burr C., Aitken M., Cowls J., Katell M., Briggs M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer. El Consejo de Europa, disponible en: https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf

422 Leslie D., Burr C., Aitken M., Cowls J., Katell M., Briggs M. (2021). Artificial intelligence, human rights, democracy, and the rule of law: a primer, The Council of Europe, disponible en: https://www.turing.ac.uk/sites/default/files/2021-03/cahai_feasibility_study_primer_final.pdf

4. Actividades

Estas actividades grupales tienen como objetivo alentar a los participantes de la capacitación a analizar y debatir casos de posibles violaciones de los derechos humanos utilizando ADM e IA en operaciones judiciales y casos de deliberación judicial de derechos humanos infringidos utilizando IA por parte de terceros.

Actividad 1

Revise el Apéndice B de la Directiva canadiense sobre la toma de decisiones automatizada y examine los cuatro niveles de impacto que una decisión asistida por IA puede tener en los derechos fundamentales.[423]

Considere el siguiente escenario: la Agencia de Empleo en el país X tiene la intención de calcular la probabilidad de que los solicitantes de empleo registrados encuentren empleo dentro de un cierto período en el futuro, teniendo en cuenta varios factores: grupo de edad de los solicitantes de empleo, género, educación, condiciones de salud, tareas de cuidado, el desempeño de su mercado laboral regional y cuánto tiempo han estado registrados en la Agencia de Empleo. En función de la probabilidad calculada, los solicitantes de empleo se asignarán a diferentes grupos: grupo uno que cubre a los solicitantes de empleo con altas oportunidades de mercado, otro grupo con oportunidades medias y un último grupo con oportunidades bajas. El sistema de IA ayudará a los asesores de las agencias de empleo a evaluar las oportunidades de los solicitantes de empleo y permitirá un uso más eficiente de los recursos. Basados en este escenario, los participantes de la capacitación examinan los cuatro niveles de impacto que una decisión tomada por el sistema de IA puede tener en los derechos de los solicitantes de empleo.[424]

Apéndice B: Niveles de evaluación de impacto	
Nivel	**Descripción**
I	Es probable que la decisión tenga poco o ningún impacto en: • los derechos de las personas o comunidades, • la salud o el bienestar de las personas o comunidades, • los intereses económicos de individuos, entidades o comunidades, • la sostenibilidad continua de un ecosistema. Las decisiones de nivel I a menudo conducirán a impactos que son reversibles y breves.
II	Es probable que la decisión tenga un impacto moderado en: • los derechos de las personas o comunidades, • la salud o el bienestar de las personas o comunidades, • los intereses económicos de individuos, entidades o comunidades, • la sostenibilidad continua de un ecosistema. Las decisiones de nivel II a menudo darán lugar a impactos que probablemente sean reversibles y a corto plazo.
III	Es probable que la decisión tenga un impacto moderado en: • los derechos de las personas o comunidades, • la salud o el bienestar de las personas o comunidades, • los intereses económicos de individuos, entidades o comunidades, • la sostenibilidad continua de un ecosistema. Las decisiones de nivel III a menudo conducirán a impactos que pueden ser difíciles de revertir y que están en curso. Probablemente se alcanzaría al menos el nivel III para las actividades policiales predictivas teniendo en cuenta el alto impacto en las libertades y los derechos de las personas y las comunidades que se han destacado anteriormente.

423 Véase: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

424 Barros Vale S., Zanfir-Fortuna G. (2022). Automated Decision-Making Under the GDPR: Practical Cases from Courts and Data Protection Authorities, disponible en: https://fpf.org/blog/fpf-report-automated-decision-making-under-the-gdpr-a-comprehensive-case-law-analysis/

Apéndice B: Niveles de evaluación de impacto	
IV	Es probable que la decisión tenga un impacto moderado en: • los derechos de las personas o comunidades, • la salud o el bienestar de las personas o comunidades, • los intereses económicos de individuos, entidades o comunidades, • la sostenibilidad continua de un ecosistema. Las decisiones de nivel IV a menudo darán lugar a impactos irreversibles y perpetuos.

Actividad 2

Plantilla de hoja informativa de evaluación de riesgos: mire la siguiente plantilla de evaluación de riesgos y vea si se le ocurre alguna otra pregunta que pueda hacer para examinar la herramienta de evaluación de riesgos.

- ¿Quién creó la evaluación de riesgos? ¿Son una organización pública o privada?
- ¿Qué tan grande era el conjunto de datos de capacitación?
- ¿Cómo se recopiló y ensambló el conjunto de datos de capacitación (es decir, de qué jurisdicción(es) proviene(n))?
- ¿En qué tiempo se recopilaron los datos?
- ¿Qué factores (es decir, las características del acusado) se incluyeron en el conjunto de datos? Esta pregunta se refiere a todos los factores que estaban disponibles sobre los acusados, no necesariamente a todos los factores que se utilizaron para capacitar o desarrollar el modelo.
- ¿El conjunto de datos incluye casos de acusados que fueron detenidos? Si es así, ¿los datos incluyen resultados para esas personas (es decir, los datos explicaron la estimación contrafactual; si es así, cómo)?
- ¿Existe algún problema o error conocido con los datos?
- ¿En qué año se creó la evaluación de riesgos?
- ¿Qué factores, entre todos los factores en los datos de capacitación, se tuvieron en cuenta en el desarrollo de la evaluación de riesgos? Si no se tuvieron en cuenta todos los factores, ¿cómo se eligieron los que sí se consideraron?
- ¿Cómo se eligieron los factores que se consideraron finalmente para la exclusión o inclusión en el modelo final (la evaluación de riesgos en sí)?
- ¿El modelo final incluye como factor(es) el arresto que no condujo a condenas? ¿El modelo final incluye factores socioeconómicos como la vivienda y la situación laboral? ¿El modelo final incluye factores de salud personal como la salud mental o el abuso de sustancias? [dividir en varias preguntas si hay información relevante disponible]
- ¿Cómo se asignó la ponderación de cada factor incluido en el modelo final? (Coeficientes de correlación de redondeo, Método de Burgess, etc.)
- ¿Cómo define el modelo final los resultados (es decir, durante el proceso de desarrollo del modelo, se definió un resultado distinto para cada tipo de falla [falla en la comparecencia, nuevo delito, nuevo delito violento, etc.] o se agravaron los resultados)?

- ¿Cómo se ve el resultado del modelo (es decir, una puntuación en una escala del 1 al 10, etc.)?
- ¿El modelo genera designaciones de nivel de riesgo o convierte las puntuaciones brutas en designaciones de nivel de riesgo como "riesgo bajo", "riesgo moderado" y "riesgo alto"?
- ¿Qué proporción de muestras en el conjunto de datos de capacitación falló en cada puntaje y/o nivel de riesgo (por ejemplo, qué porcentaje de personas con un puntaje de 5 o una etiqueta de "riesgo moderado" en realidad no compareció)?
- ¿Los desarrolladores del modelo evaluaron la validez predictiva del mismo? En caso afirmativo, ¿cómo?
- ¿Dónde se utiliza la evaluación de riesgos?
- ¿Los factores y las ponderaciones de la evaluación de riesgos están disponibles públicamente?
- ¿La evaluación de riesgos cuesta dinero para que una jurisdicción la adopte?
- ¿La adopción de la evaluación de riesgos requiere capacitación? Si es así, ¿por quién?
- ¿La evaluación de riesgos incluye algún tipo de software o paquete de software?
- ¿La evaluación de riesgos implica o requiere una entrevista en persona?
- ¿Cómo explica la evaluación de riesgos la información faltante?
- ¿Se ha analizado la evaluación de riesgos en datos no relacionados con la capacitación para determinar la validez predictiva? ¿Se ha analizado la evaluación de riesgos con datos de capacitación o datos no relacionados con esta sobre el rendimiento para diferentes grupos de razas? ¿Se ha analizado la evaluación de riesgos con datos de capacitación o datos no relacionados con esta sobre el rendimiento para diferentes grupos de géneros? Si es así, ¿por quién, cuándo y con qué datos?[425]

Actividad 3

Debata las siguientes preguntas con otros participantes de la capacitación:

- ¿Qué implica la privacidad en una era en la que la recopilación de datos en tiempo real es común y existe la posibilidad de violaciones de datos, robo de identidad o fraude en línea?
- ¿Podemos expresarnos libremente en todas las herramientas y plataformas digitales sin preocuparnos por la censura o la elaboración de perfiles de IA?
- ¿Puede todo el mundo tener el mismo acceso a información fiable dada la difusión generalizada de material dañino y mentiras en línea?
- ¿Cómo podemos garantizar que las tecnologías de IA ayuden a cerrar la brecha digital en lugar de ampliar las disparidades ya existentes?

425 Véase: https://law.stanford.edu/pretrial-risk-assessment-tools-factsheet-project/

5. Recursos

1. Access Now (2018). Mapping artificial intelligence strategies in Europe: a new report by Access Now, disponible en: https://www.accessnow.org/mapping-artificial-intelligence-strategies-in-europe/
2. Artículo 19, Instituto Danés de Derechos Humanos (2017). Ejemplo de herramienta de evaluación de impacto de derechos humanos de ccTLD, disponible en: https://www.article19.org/wp-content/uploads/2017/12/Sample-ccTLD-HRIA-Dec-2017.pdf
3. Algoritmos de auditoría: adición de responsabilidad a la autoridad automatizada, disponible en: http://auditingalgorithms.science/
4. Comisión Australiana de Derechos Humanos (2018). Informe final: Derechos humanos y tecnología, disponible en: https://tech.humanrights.gov.au/sites/default/files/2018-07/Human%20Rights%20and%20Technology%20Issues%20Paper%20FINAL.pdf
5. CAHAI (2020). Marco jurídico de los sistemas de IA. Estudio de viabilidad de un marco legal para el desarrollo, diseño y aplicación de la inteligencia artificial, basado en los estándares del Consejo de Europa sobre derechos humanos, democracia y Estado de derecho, Estudio del Consejo de Europa, DGI/2021/04, disponible en: https://edoc.coe.int/en/artificial-intelligence/9648-a-legal-framework-for-ai-systems.html
6. Consejo Europeo (2020). Recomendación CM/Rec (2020) del Comité de Ministros a los Estados miembros sobre los impactos de los sistemas algorítmicos en los derechos humanos, disponible en: https://rm.coe.int/09000016809e1154
7. Dearden L. (2018). La nueva tecnología puede detectar videos de ISIS antes de que se carguen, disponible en: http://www.independent.co.uk/news/uk/home-news/isis-videos-artificial-intelligence-propaganda-ai-home-office-islamic-state-radicalisation-asi-data-a8207246.html
8. Duarte N., Llanso E., Loup A. (2017), Mixed Messages? Los límites del análisis automatizado del contenido de las redes sociales, disponible en: https://cdt.org/insight/mixed-messages-the-limits-of-automated-social-media-content-analysis/
9. Elsayed-Ali S. (2017). Inteligencia artificial y el futuro de los derechos humanos, disponible en: https://medium.com/amnesty-insights/artificial-intelligence-and-the-future-of-human-rights-b58996964df5
10. Edwards L. (2022). Opinión de expertos: Regulación de la IA en Europa. Cuatro problemas y cuatro soluciones, disponible en: https://www.adalovelaceinstitute.org/report/regulating-ai-in-europe/
11. EUR-Lex (2021). Proyecto de Reglamento de IA de la UE, disponible en : https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
12. Comisión Europea (2021). Estudio para apoyar una evaluación de impacto de los requisitos reglamentarios para la inteligencia artificial en Europa, disponible en: https://artificialintelligenceact.eu/wp-content/uploads/2022/06/AIA-COM-Impact-Assessment-3-21-April.pdf
13. Supervisor Europeo de Protección de Datos. Necesidad y proporcionalidad, disponible en: https://edps.europa.eu/data-protection/our-work/subjects/necessity-proportionality_en
14. ICO, Kit de herramientas sobre IA y riesgos para la protección de datos, disponible en: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/
15. Jones K. (2023), disponible en: https://www.chathamhouse.org/2023/01/ai-governance-and-human-rights

16. Latonero M. (2018). Artificial Intelligence & Human Rights: A Workshop at Data & Society, disponible en: https://points.datasociety.net/artificial-intelligence-human-rights-a-workshop-at-data-society-fd6358d72149
17. Latonero M. (2019). Governing Artificial Intelligence: Upholding Human Rights and Human Dignity, disponible en: https://datasociety.net/wpcontent/uploads/2018/10/DataSociety_Governing_Artificial_Intelligence_Upholding_Human_Rights.pdf
18. L.iberty (2020). Liberty wins ground-breaking victory against facial recognition tech, disponible en: https://www.libertyhumanrights.org.uk/issue/liberty-wins-ground-breaking-victory-against-facial-recognition-tech/
19. Mounk Y. (2018). Verboten. Germany's risky law for stopping hate speech on Facebook and Twitter, disponible en: https://newrepublic.com/article/147364/verboten-germany-law-stopping-hate-speech-facebook-twitter
20. Observatorio de Políticas de IA de la OCDE. Datos en tiempo real, disponible en: https://oecd.ai/en/data?selectedArea=ai-research&selectedVisualization=top-countries-in-ai-scientific-publications-in-time-from-scopus
21. Observatorio de Políticas de IA de la OCDE. Descripción general de los principios, disponible en: https://oecd.ai/en/ai-principles
22. Ortiz Freuler J., Iglesias C. (2018). Algorithms and Artificial Intelligence in Latin America: A Study of Implementation by Governments in Argentina and Uruguay, World Wide Web Foundation, disponible en: http://webfoundation.org/docs/2018/09/WF_AI-in-LA_Report_Screen_AW.pdf
23. Peralta Gutiérrez (2022). Marco normativo de la Inteligencia Artificial en el ámbito comparado. In: Herrera Triguero F., Peralta Gutiérrez A., Torres López L.S., El derecho y la inteligencia artificial, EUG: Granada 189–222.
24. Pielemeier J. (2018). The Advantages and Limitations of Applying the International Human Rights Framework to Artificial Intelligence, disponible en: https://points.datasociety.net/the-advantages-and-limitations-of-applying-the-international-human-rights-framework-to-artificial-291a2dfe1d8a
25. Reiling D., Contini F. (2022). E-Justice Platforms: Challenges for Judicial Governance, International Journal for Court Administration, 13 (1), disponible en: https://iacajournal.org/articles/10.36745/ijca.445
26. Reisman D., Schultz J., Crawford K., Whittaker M. (2018). Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability, disponible en: https://ainowinstitute.org/publication/algorithmic-impact-assessments-report-2
27. Reitman R. (2022). Podcast Episode: Algorithms for a Just Future, disponible en: https://www.eff.org/deeplinks/2022/01/podcast-episode-algorithms-just-future
28. Stankovich M. (2021). Regulating AI and Big Data Deployment in Healthcare: Proposing Robust and Sustainable Solutions for Developing Countries 'Governments, disponible en: https://www.dai.com/uploads/regulating-ai-cda.pdf
29. Vincent J. (2019). AI won't relieve the misery of Facebook's human moderators, disponible en: https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms
30. YouTubeHelp. How Content ID Works, disponible en: https://support.google.com/youtube/answer/2797370?hl=en

Recursos sugeridos por la UNESCO

Materiales UNESCO

Publicaciones

Caja de Herramientas Global para Actores Judiciales: Normas jurídicas internacionales sobre libertad de expresión, acceso a la información y seguridad de los periodistas

- ☐ Disponible en: árabe, chino, español, francés, inglés, portugués y ruso

Caja de herramientas para el Poder Judicial en África sobre los estándares legales relativos a la libertad de expresión

- ☐ Disponible en: inglés, francés y portugués

Directrices para los fiscales sobre casos de delitos contra periodistas

- ☐ Disponible en: amárico; árabe; chino; darí; inglés; francés; indonesio; italiano; jemer; portugués; pastún; ruso; somalí; español; suajili; tailandés; ucraniano; y uzbeko

Directrices para los actores judiciales sobre privacidad y protección de datos

- ☐ Disponible en: árabe, chino, español, francés, inglés, portugués y ruso

COVID-19: Directrices sobre el papel de los operadores judiciales en la protección y promoción del derecho a la libertad de expresión

- ☐ Disponible en: árabe; birmano; chino; inglés; francés; jemer; portugués; ruso; y español

Seguridad de los periodistas que cubren las protestas: preservar la libertad de prensa en tiempos de agitación

- ☐ Disponible en: árabe , chino, español, francés, inglés, portugués y ruso

Caja de herramientas global para agentes de la ley: libertad de expresión, acceso a la información y seguridad de los periodistas

- ☐ Disponible en: árabe; inglés; francés; español; chino; portugués y ruso

Folleto sobre Libertad de expresión y orden público: fomentar la relación entre las fuerzas de seguridad y los periodistas

- ☐ Disponible en: inglés, portugués, ruso, ucraniano y somalí

El “mal uso” del sistema judicial para atacar la libertad de expresión: tendencias, desafíos y respuestas

- ☐ Disponible en: árabe, chino, inglés, francés, italiano, portugués, ruso, y español

Guía de la UNESCO para intervenciones amicus curiae en casos de libertad de expresión

- ☐ Disponible en: árabe, chino, inglés, francés, ruso y español

Series de videos y seminarios web

La próxima frontera: la propiedad intelectual en la era de la inteligencia artificial generativa

- ☐ Disponible en: inglés y español

El desafío de admisibilidad: evidencia generada por IA en la sala del tribunal

- ☐ Disponible en: inglés

Foro de Gobernanza de Internet 2021: La inteligencia artificial y el Estado de derecho en el ecosistema digital

- ☐ Disponible en: inglés

Foro de Gobernanza de Internet 2022: ¿Por qué la transformación digital y la inteligencia artificial son importantes para la justicia?

- ☐ Disponible en: inglés

Explicadores en vídeo de la UNESCO: Cómo poner fin a la impunidad de los crímenes contra periodistas

- ☐ Disponible en: árabe, chino, inglés, francés, ruso y español

La prueba de las tres partes: límites legítimos a la libertad de expresión

- ☐ Disponible en: árabe, chino, inglés, francés, portugués, ruso y español

El Plan de acción de Rabat sobre la prohibición de la incitación al odio: límites legítimos a la libertad de expresión

- ☐ Disponible en: árabe, chino, inglés, francés, portugués, ruso y español

Explicaciones en vídeo de la UNESCO: ¿Cómo sería un mundo sin medios de comunicaciónindependientes?

- ☐ Disponible en: árabe, chino, inglés, francés, ruso y español

Explicadores en video de la UNESCO: ¿Por qué #LibertadDeExpresión y #AccesoALaInformación son tan importantes para las elecciones libres y justas?

☐ Disponible en: árabe, chino, español, francés, inglés, portugués y ruso

Explicadores en video de la UNESCO: Tribunales judiciales regionales en África y jurisprudencia histórica sobre la libertad de expresión

☐ Disponible en: inglés, francés y portugués

Desafíos legales relacionados con la libertad de expresión en medio de la pandemia de COVID-19

☐ Disponible en: inglés, francés y español

Cursos

Curso masivo abierto en línea (MOOC) sobre inteligencia artificial y Estado de derecho

☐ Disponible en: árabe, chino, inglés, francés, portugués, ruso y español

Nuevo Curso multilingüe abierto en línea (MOOC) del Instituto Bonavero-UNESCO sobre libertad de expresión y seguridad de los periodistas

☐ Disponible en: inglés, árabe, chino, francés, portugués, ruso y español

¿Cómo hacer uso de este Kit de herramientas?

Este Kit de herramientas fomenta un modelo pedagógico experiencial: no pretende ser prescriptivo, y se alienta a las personas usuarias a aprovechar sus propias experiencias teniendo en cuenta los contextos relevantes en los que se utiliza el Kit de herramientas. Aunque está dirigido principalmente a operadores judiciales, también puede ser útil para una variedad de terceros, incluidas las organizaciones de la sociedad civil. Hay varias formas diferentes en que el kit de herramientas se puede utilizar como recurso:

- Taller presencial integral: aconsejamos que un taller presencial integral que cubra los cuatro módulos tenga una duración de al menos tres días. En circunstancias en las que los participantes no estén familiarizados con los principios fundamentales del derecho internacional de los derechos humanos, aconsejamos que el taller se lleve a cabo durante al menos cuatro días.
- Taller específico: también se podrían realizar talleres sobre módulos seleccionados dentro del Kit de herramientas. En tales circunstancias, los capacitadores deben asegurarse de que se establezcan las bases de los otros módulos que puedan ser necesarios para que los participantes comprendan completamente los conceptos y lleven a cabo las tareas.
- Curso en línea combinado (como un curso en línea abierto masivo) y taller presencial: este formato proporcionaría más tiempo para que los participantes tengan contacto con los materiales y los ejercicios de autoevaluación, antes de reunirse en el entorno presencial. Idealmente, el componente en línea debería estar respaldado por foros de discusión en línea y otro tipo de apoyo.
- Autoaprendizaje: el kit de herramientas es de naturaleza autoexplicativa y puede servir como un recurso de autoaprendizaje útil para participar individualmente o entre un grupo de personas que trabajan en una organización en particular. Si bien a menudo es beneficioso que haya debates colaborativos y el intercambio de experiencias, también puede ser un punto de partida útil y una referencia para alguien que busca aumentar su comprensión de los problemas emergentes en IA y derechos humanos.

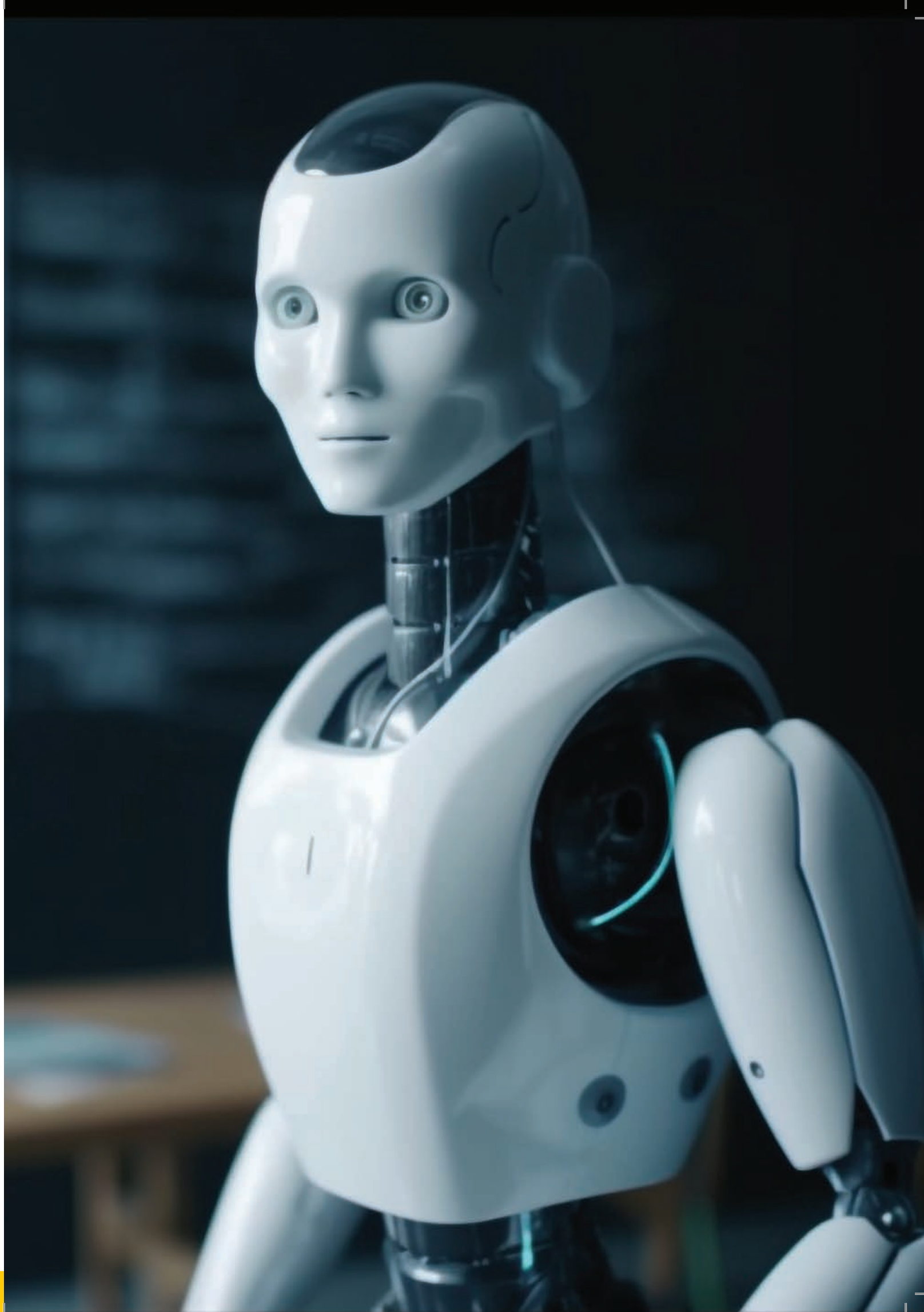
Aunque los talleres más largos permitirán realizar más actividades, es poco probable que haya tiempo para realizar todas las actividades sugeridas. Esto queda a discreción de los capacitadores. Los capacitadores deben tratar de evaluar los aspectos que son más relevantes para los participantes y cuál es la mejor forma de integrarlos en sus escenarios laborales y locales.

Se recomienda que los capacitadores distribuyan un cuestionario a los participantes previamente para determinar su experiencia en esta área de la ley. La siguiente plantilla podría adaptarse en función de los participantes esperados en la capacitación:

Kit global de herramientas globaloperadores judiciales sobre IA y Estado de derecho

Datos del participante **Nombre:** **Organización:** **Designación:** **País**	**Experiencia del participante** **¿Tiene experiencia jurídica?** **¿Tiene experiencia en la toma de decisiones automatizada, la IA y los derechos humanos?** **Brinde información más detallada.**
¿Qué módulos serían de mayor utilidad para usted y su trabajo? Seleccione. ☐ **Módulo 1. Introducción a la IA y el Estado de derecho** ☐ **Módulo 2. Adopción de IA en el poder judicial** ☐ **Módulo 3. Desafíos legales y éticos del despliegue de IA en el poder judicial** ☐ **Módulo 4. Derechos humanos e IA: gobernanza, regulación y política**	
Brinde información más detallada.	**¿Cuáles son sus objetivos para esta capacitación?**

La "IA y los derechos humanos" es un área de la ley dinámica y en evolución. Como tal, es probable que haya nuevos desarrollos frecuentes. Los instructores deben mantenerse al tanto de estos desarrollos y actualizar el material de capacitación en consecuencia.



Anexo I

Evaluación del impacto ético de la UNESCO para los sistemas de IA

Este instrumento tiene dos objetivos: en primer lugar, evaluar si los algoritmos específicos están alineados con los valores, principios y directrices establecidos por la Recomendación. En segundo lugar, garantizar la transparencia pidiendo que la información sobre los sistemas de IA y la forma en que se desarrollaron estén disponibles para el público. Así no es como funciona hoy en día, incluso para obtener información básica sobre la seguridad y la fiabilidad de la IA.

Las herramientas de evaluación de impacto están ganando terreno para evaluar el verdadero impacto de los sistemas de IA. De hecho, las evaluaciones de impacto son obligatorias por el proyecto de ley de IA de la UE para los sistemas de alto riesgo, y se proponen como parte del debate del Consejo de Europa sobre una Convención para la IA.

La Recomendación de la UNESCO es única, ya que considera todo el ciclo de vida de la IA. Por lo tanto, la Evaluación de impacto ético incluye requisitos ex ante y ex post. En una etapa inicial, establece la importancia de asegurar los datos de calidad y representatividad, la diversidad de los equipos que desarrollan los productos, la robustez y transparencia de los algoritmos, su auditabilidad y la posibilidad de insertar puntos de control en diferentes momentos del proceso de desarrollo.

La EIA se propone a los compradores de sistemas de IA, ya que este es uno de los principales canales en los que los algoritmos se abren paso a dominios públicos altamente sensibles. Pero las preguntas y la estructura del documento están diseñadas para que las herramientas también puedan utilizarse de manera más general por los desarrolladores de sistemas de IA, en los sectores público o privado, que deseen desarrollar la IA de manera ética y cumplir plenamente con los estándares internacionales como la Recomendación.

El documento consta de dos partes principales que en conjunto logran un equilibrio entre el procedimiento y el fondo. En la primera parte, relacionada con el alcance, el objetivo es comprender los conceptos básicos del sistema, así como plantear algunas preguntas preliminares, como si la automatización es la mejor solución para el caso en cuestión.

También plantea preguntas sobre el equipo del proyecto y si existen planes para involucrar a diferentes partes interesadas. La segunda parte está dedicada a la aplicación de los principios de la Recomendación de la UNESCO.

Para cada principio, las preguntas tendrán como objetivo evaluar:

a. Si se han establecido suficientes garantías procesales para garantizar que el sistema cumpla con la Recomendación; y

b. Los resultados positivos (potenciales) y los impactos adversos que pueden surgir de la adquisición y el despliegue del sistema, específicos de su contexto de uso.

La herramienta de evaluación está disponible en: https://unesdoc.unesco.org/ark:/48223/pf0000386276

Anexo II

Ejemplos de actividades adicionales

- Estudios de caso de sistemas de IA en servicios públicos en América Latina - http://webfoundation.org/docs/2018/09/WF_AI-in-LA_Report_Screen_AW.pdf
- Actividad interactiva (juego de algoritmos de la corte) de usar Compas - https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/
- Manual de IA confiable: https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf
- Actividad de evaluación de impacto algorítmico, Gobierno de Canadá, Algorithmic Impact Assessment. https://canada-ca.github.io/aia-eia-js/
- Lista de evaluación para el ejercicio de IA confiable (ALTAI)

Herramienta de mapeo de IA[426]

#	Pregunta	Respuestas
1	¿Cuál es el nombre de la herramienta de Inteligencia Artificial que se está evaluando con este cuestionario?	
2	Describa brevemente la funcionalidad principal de la herramienta.	
3	¿Qué motiva el uso de herramientas de IA en este caso? (Marque todas las respuestas que correspondan)	1) Trabajo o casos atrasados existentes 2) Mejorar la calidad general de las decisiones 3) Menores costos de transacción de un programa existente 4) La herramienta está realizando tareas que los humanos no podrían realizar en un período de tiempo razonable 5) Uso de métodos innovadores 6) Otros
4	¿Cómo se desarrolló esta herramienta?	1) Fue completamente desarrollada por el personal técnico de su institución 2) Fue desarrollada en colaboración con una entidad externa. 3) Fue adquirida y/o desarrollada en su totalidad por una parte externa 4) No lo sé 5) Otros:

426 Brehm K., Hirabayashi M., Langevin C., Munoscano B.R., Sekizawa K., Shu J. (2020). El futuro de la IA en el sistema judicial brasileño: mapeo, integración y gobernanza de la IA. El futuro de la IA en el sistema judicial brasileño. AI Mapping, Integration, and Governance, Technical report, ITS Rio, disponible en: https://itsrio.org/wp-content/uploads/2020/06/SIPA-Capstone-The-Future-of-AI-in-the-Brazilian-Judicial-System-1.pdf

#	Pregunta	Respuestas
5	¿Para qué plataforma de Justicia electrónica se ha desarrollado esta herramienta?	
6	¿En qué etapa de desarrollo se encuentra actualmente la herramienta?	1) En desarrollo/proceso de adquisición en curso 2) Como prototipo/En prueba 3) Lista para la implementación, no está operativa actualmente 4) Totalmente implementada 5) Otros:
7	¿En qué métodos se basa la herramienta?	1) Regresión logística 2) Máquinas de vectores de soporte 3) Árboles de decisión/Bosque aleatorio 4) Redes neuronales/CNN 5) Métodos de sobremuestreo/re-muestreo 6) Métodos de reducción de la dimensionalidad (PCA, agrupación, aprendizaje múltiple) 7) Otros:
8	Marque cuál de las siguientes capacidades, si corresponde, se aplica a la herramienta. (Marque todas las respuestas que correspondan)	1) Modelado y evaluación de riesgos: análisis de conjuntos de datos para identificar patrones y recomendar cursos de acción y, en algunos casos, desencadenar acciones específicas. 2) Organización de datos: análisis de datos para categorizar, procesar, clasificar, personalizar y ofrecer contenido específico para contextos específicos. 3) Reconocimiento de imágenes y objetos: análisis de datos para automatizar el reconocimiento, la clasificación y el contexto asociados con una imagen u objeto. 4) Análisis de texto y habla: analizar datos para reconocer, procesar y etiquetar texto, habla, voz y hacer recomendaciones, clasificaciones u otro tipo de resultados basados en el etiquetado. 5) Optimización de procesos y automatización de flujos de trabajo: análisis de datos para identificar anomalías, patrones de clúster, predecir resultados o formas de optimizar; y automatizar flujos de trabajo específicos. 6) Ninguno/No aplica 7) Otros
9	¿La herramienta realiza algún tipo de análisis de datos no estructurados?	1) Sí 2) No 3) I don't know. 4) No lo sé.

#	Pregunta	Respuestas
10	¿El equipo que la utiliza conoce los datos que se utilizaron para capacitar a la herramienta?	1) Sí 2) No 3) No lo sé
11	¿El código de la herramienta está disponible públicamente y es revisable?	1) Sí 2) No 3) No lo sé. 4) No se aplica.
12	El algoritmo de la herramienta y su código son:	1) Código abierto 2) Propiedad del tribunal 3) Propiedad de un tercero
13	¿La herramienta recopila y/o analiza datos personales (según lo define la Ley General de Protección de Datos)?	1) Recopila 2) Analiza 3) Ninguno.
14	¿La herramienta recopila y/o analiza información de identificación personal?	1) Recopila 2) Analiza 3) Ninguno
15	Los datos utilizados por la herramienta... (Marque todas las respuestas que correspondan)	1) Fueron recopilados por un tribunal o una entidad gubernamental 2) Están disponibles públicamente y son revisables 3) Se comparten con otra entidad. 4) Fueron recopilados por una entidad externa. 5) Se comparten con una entidad externa.
16	El personal técnico de su institución está en capacidad de explicar:	1) Cuáles son las entradas de la herramienta 2) Cuáles son las salidas de la herramienta. 3) El proceso a través del cual las entradas se convierten en salidas.
17	El personal no técnico de su institución está en capacidad de explicar:	1) Cuáles son las entradas de la herramienta. 2) Cuáles son las salidas de la herramienta. 3) El proceso a través del cual las entradas se convierten en salidas.
18	La herramienta ha pasado por:	1) Un seguimiento técnico y procesos de aseguramiento de la calidad 2) Una revisión de sus datos de capacitación para detectar sesgos 3) Una revisión jurídica y/o administrativa 4) Otros:

Anexo III

Agenda de capacitación – plantilla

Kit de herramientas globales sobre IA y el Estado de derecho para el poder judicial

Capacitación de 3 días

Fecha:

Título	Capacitación para [insertar público objetivo] sobre el Kit de herramientas de capacitación mundial de la UNESCO sobre IA y el Estado de derecho para el poder judicial
Modalidad	Física
Publico meta/objetivo	
Fechas	
Duración	3 días
Descripción	El programa de capacitación se basa en el Kit de herramientas mundial sobre la IA y el Estado de derecho para el Poder judicial
Organización	La capacitación será organizada por la UNESCO
Ubicación	
Plazo de inscripción	
Costos de la capacitación	
Idioma	

1. OBJETIVOS DEL APRENDIZAJE

Este programa de capacitación tiene como objetivo proporcionar a los operadores judiciales acceso a la información y las herramientas necesarias para comprender y considerar los beneficios de la Inteligencia artificial ("IA") para sus operaciones. Al mismo tiempo, el programa de capacitación ayudará al poder judicial a reconocer los inconvenientes y riesgos de la IA, incluidos los prejuicios, la discriminación, las cajas negras y la falta de responsabilidad y transparencia. El programa de capacitación ayudará a los operadores judiciales a tomar mejores decisiones y reducir los posibles riesgos para los derechos humanos al ofrecer orientación y perspectivas sobre los principios, las regulaciones y la jurisprudencia relevante que sustentan el uso responsable de la IA en contextos judiciales, y en general.

Para equilibrar las oportunidades y los desafíos que las tecnologías de IA pueden presentar para el sector de la justicia, la Recomendación de la UNESCO sobre la Ética de la IA destaca que "los Estados miembros deben mejorar la capacidad del poder judicial para tomar decisiones relacionadas con los sistemas de IA según el Estado de derecho...". De ahí la importancia de este programa de capacitación para establecer cómo el sector de la justicia puede aprovechar las tecnologías de IA y garantizar que se utilicen de manera ética, responsable y de acuerdo con el marco del derecho internacional de los derechos humanos.

2. RESULTADOS DEL APRENDIZAJE

Después de completar el programa de capacitación, los operadores judiciales podrán:

- Comprender la IA y la toma de decisiones algorítmicas (ADM) y su uso en los procesos y operaciones judiciales.
- Entender que la IA no es neutra, es un sistema sociotécnico que representa el mundo que nos rodea.
- Desarrollar la capacidad de examinar casos judiciales relacionados con el uso de la IA.
- Comprender las cuestiones fundamentales relacionadas con el sesgo algorítmico (como el sesgo de género, el sesgo racial, las formas de sesgo que se cruzan, etc.) y las cajas negras y explicar por qué son importantes en los entornos judiciales.
- Familiarizarse con las medidas regulatorias y la jurisprudencia más recientes relacionadas con el sesgo algorítmico, el uso inapropiado de algoritmos en la toma de decisiones, incluso en contravención de la ley, y las cajas negras.
- Entender y explicar el impacto de la IA en los siguientes derechos fundamentales: privacidad, libertad de expresión, acceso a la información, protección contra la discriminación, derecho al acceso a los tribunales, juicios y audiencias justos e imparciales y debido proceso judicial.

3. PÚBLICO OBJETIVO

El público objetivo principal de la capacitación se compone de operadores judiciales, centrándose principalmente en los jueces. La capacitación

también puede incluir fiscales, procuradores, abogados públicos, otras partes interesadas del sector de la justicia en todo el mundo y empresas de tecnología legal.

4.REQUISITOS DE ACCESO

Kit de herramientas globales sobre IA y el Estado de derecho para el Poder judicial

5. INSTRUCTORES

NOMBRE DE LOS INSTRUCTORES	DATOS DE CONTACTO

6. CONTENIDOS DEL CURSO DE CAPACITACIÓN

El programa de capacitación se basa principalmente en el Kit de herramientas mundial sobre la IA y el Estado de derecho para el poder judicial y abarcará los siguientes temas:

1. Módulo 1: Introducción a la IA y al Estado de derecho
2. Módulo 2: Adopción de IA en el poder judicial
3. Módulo 3: Desafíos legales y éticos de la IA
4. Módulo 4: Derechos humanos e IA

7. CONTENIDOS Y AGENDA DEL PROGRAMA DE CAPACITACIÓN

Día 1: Introducción a la IA y su uso en el poder judicial

Hora	Orden del día
8:30 – 9:00	Inicio de sesión y registro del participante
9:00 – 9:30	Apertura e introducción a los objetivos del programa de capacitación
9:30 – 11:00	Sesión 1: Comprender la IA y sus componentes básicos Moderador/a: Esta sesión tiene como objetivo proporcionar una comprensión integral de la IA mediante la exploración de su definición y los componentes básicos clave. A través de debates atractivos, ejemplos ilustrativos, estudios de casos y actividades grupales, examinaremos los diversos componentes de los sistemas de IA, incluidos los algoritmos, el aprendizaje automático, los datos y los modelos computacionales. Esta sesión también abordará los riesgos fundamentales relacionados con el desarrollo y la implementación de la IA, como el sesgo, las cajas negras y la ciberseguridad. Al final de esta sesión, los participantes obtendrán una sólida comprensión de los conceptos fundamentales relacionados con la IA, lo que les permitirá desenvolverse en este campo con confianza y claridad.
11:00 – 11:30	Pausa para café

11:30 – 13:00	Sesión 2: ¿Cuáles son los usos de la IA en el sector de la justicia? Facilitator: Esta sesión describirá algunos usos principales de la IA en el poder judicial, como el descubrimiento electrónico y la revisión de documentos, el uso de IA generativa para ayudar con la redacción de documentos, el análisis predictivo y el soporte de ADM, las herramientas de evaluación de riesgos, la resolución de disputas, el reconocimiento y análisis de idiomas, el archivo digital y la gestión de casos.
13:00 – 14:30	Almuerzo
14:30 – 16:00	Sesión 3: Estudios de caso sobre el uso de IA en el poder judicial Moderador/a: Esta sesión examinará estudios de casos seleccionados sobre el despliegue de IA en el sistema de justicia, como VICTOR, Brasil, el Sistema Inteligente de Transcripción de Tribunales de Singapur, Prometea, Argentina, PretorIA, Colombia, el uso de IA en el sistema de justicia de China, el uso de IA en el sistema de justicia de la India, HART (Herramienta de Evaluación de Riesgos de Daños) del Reino Unido, PredPol y Palantir. La sesión invitará a los participantes a compartir su experiencia con los sistemas de IA y facilitará una conversación más amplia sobre las oportunidades, los desafíos y los riesgos asociados con el uso de estos sistemas en el poder judicial.
16:00 – 16:30	Comentarios y evaluación
16:30 – 16:45	Conclusión del primer día y esbozo del orden del día del segundo día

Día 2: Cuestiones legales y éticas relacionadas con los sistemas de IA

Hora	Orden del día
8:30 – 9:00	Inicio de sesión y registro del participante
9:00 – 11:00	Sesión 4: Responsabilidad algorítmica y transparencia Moderador/a: A través de debates perspicaces y estudios de casos del mundo real, esta sesión guiará a los participantes a través de los conceptos de transparencia algorítmica y conceptos de rendición de cuentas y destacará los problemas legales fundamentales que los operadores judiciales deben conocer. Se prestará especial atención a la identificación biométrica, el reconocimiento facial y las falsedades profundas.
11:00 – 11:30	Pausa para café

11:30 – 13:00	Sesión 5: Jurisprudencia emergente sobre sesgos y cajas negras Moderador/a: La sesión presentará la jurisprudencia existente que se ocupa de las cajas negras algorítmicas y el sesgo en los ADM y los sistemas de IA utilizados en la prestación de servicios públicos y por el sector privado. Utilizando estudios de casos de la vida real, los participantes analizarán cómo el sesgo y las cajas negras han resultado en la violación de los derechos humanos o cualquier otro daño, y cómo los tribunales en diferentes jurisdicciones se han enfrentado a ello. Los participantes debatirán los problemas de responsabilidad por el daño causado por estos sistemas, así como el uso de la IA con fines probatorios. Los casos analizados incluirán: Caso Deliveroo (2021), Caso Foodinho (2021), El pueblo vs. Chubbs (2015), Estado de Nueva Jersey vs. Francisco Arteaga, Estado vs. Loomis, El pueblo vs. Alvin Davis, Estado de Nueva Jersey vs Pickett, caso Uber relativo al uso del programa de detección de fraudes Mastermind Estados Unidos vs Ellis y el caso australiano de Robodebt.
13:00 – 14:30	Almuerzo
14:30 – 16:00	Sesión 6: Evaluación del impacto ético de los sistemas de IA Moderador/a: Esta sesión presentará a los participantes cuestiones fundamentales relacionadas con la ética de la IA, así como los marcos éticos clave de la IA a nivel internacional, regional y nacional. Utilizando la Evaluación del impacto ético de los sistemas de IA de la UNESCO, los participantes evaluarán escenarios hipotéticos como parte de grupos de trabajo.
16:00 – 16:30	Comentarios y evaluación
16:30 – 16:45	Conclusión del primer día y esbozo del orden del día del segundo día

Día 3: IA y derechos humanos

Hora	Orden del día
8:30 – 9:00	Inicio de sesión y registro del participante
9:00 – 11:00	Sesión 7: Derechos humanos e IA: derecho al acceso a los tribunales, a un juicio justo y al debido proceso, a un remedio efectivo y a la protección contra la discriminación. Moderador/a: Numerosas aplicaciones de la IA tienen el potencial de afectar directamente la igualdad de acceso a los derechos fundamentales, incluido el derecho a la privacidad y la protección de la información personal, el derecho al acceso a la justicia y el derecho a un juicio justo, particularmente en lo que respecta a la presunción de inocencia y la carga de la prueba, el derecho al empleo, la educación, la vivienda y la salud, así como el derecho a los servicios públicos y el bienestar. Si no van acompañadas de salvaguardias adecuadas contra los prejuicios, las tecnologías de IA podrían contribuir a negar el acceso a los derechos que afectan de manera desproporcionada a las mujeres, las minorías y las personas que ya son las más vulnerables y marginadas.
11:00 – 11:30	Pausa para café

11:30 – 13:00	• Sesión 8: Derechos humanos e IA: (i) libertad de expresión, (ii) derecho a la privacidad y protección de datos, y (ii) acceso a la información. Moderador/a: Esta sesión presentará y analizará algunos de los derechos humanos afectados por los sistemas de IA implementados por terceros y adjudicados por los tribunales, como la libertad de expresión, el derecho a la privacidad y la protección de datos, y el acceso a la información. La sesión también analizará la jurisprudencia pertinente relacionada con los derechos humanos y las aplicaciones de IA.
13:00 – 14:30	Almuerzo
14:30 – 16:00	Sesión 9: Problemas emergentes en la intersección de la IA y el derecho Moderador/a: - La sesión analizará brevemente las preocupaciones en torno a: - Ciberseguridad - Derechos de propiedad intelectuals - Pruebas generadas por IA en los tribunales - Uso de RA y RV en los tribunales
16:00 – 16:30	Comentarios y evaluación
16:30 – 17:00	Resumen y conclusión del programa

8. METODOLOGÍA (enfoque didáctico)

El programa de capacitación se basa en el Kit de herramientas mundial sobre la IA y el Estado de derecho para el poder judicial. El kit de herramientas incluye actividades y recursos pertinentes a la IA, los derechos humanos y el Estado de derecho para los operadores judiciales.

Esta capacitación se impartirá físicamente e incluirá conferencias, ejercicios interactivos y debates. La capacitación se impartirá utilizando diapositivas de PowerPoint, materiales de referencia seleccionados y cuestionarios diarios de autoevaluación. Los participantes deben revisar, estudiar, participar en las actividades programadas y realizar autoevaluaciones.

9. EVALUACIÓN Y CALIFICACIÓN

El desempeño de los participantes en esta capacitación se determinará utilizando una combinación de calificaciones para los debates de las sesiones de participación y los cuestionarios de autoevaluación.

- La participación en las sesiones se valorará con un 30 por ciento.
- Los cuestionarios de autoevaluación valdrán el 70 por ciento de la calificación final de la capacitación. Habrá 6 preguntas por cuestionario.

Al final, los participantes recibirán un certificado de finalización

10. PRESENTACIONES DE CAPACITACIÓN

Se pueden utilizar una serie de presentaciones para cada módulo y en diferentes idiomas para las capacitaciones disponible en: https://unesco.sharepoint.com/:f:/s/InnovationDigitalTransformationTeam/Et10B3ujh7NMmI8jIqx_cm0BHefod3iBWzLI-w9BcTJd_A?e=SSbXh8