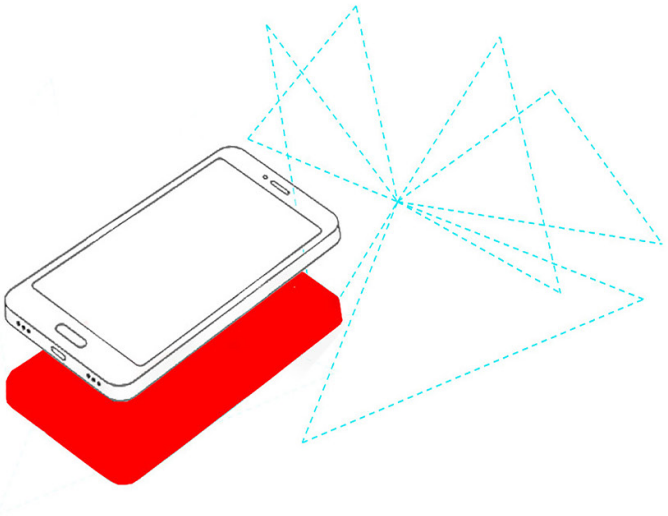
derecho digital_



CONECTADAS Y DESIGUALES: LA BRECHA DE GÉNERO EN LA ERA DIGITAL

Director

Miguel Ángel Sánchez Huete

Profesor de Derecho Financiero y Tributario

Universitat Autònoma de Barcelona

Yolanda García Calvente
Consuelo Ruiz de la Fuente
José Antonio Fernández Amor
Carolina Gala Durán
Zuley Fernández Caballero
Josep Cañabate Pérez
Mercedes Ruiz Garijo
Susana Navas Navarro
Miguel Angel Sánchez Huete
Paz Lloria García
Elisa García Mingo
Jacinto G. Lorca
Isabel Carrillo Flores
Pilar Prat Viñolas
Arantza Libano Beristain
Noelia Igareda González

DONES I DRETS
CENTRE D'ESTUDIS I DE RECERCA

REUS
EDITORIAL

UAB
Universitat Autònoma de Barcelona

COLECCIÓN DERECHO DIGITAL

TÍTULOS PUBLICADOS

1. **Datos sanitarios electrónicos. El espacio europeo de datos sanitarios**, Susana Navas Navarro, (2023).
2. **Perspectivas regulatorias de la inteligencia artificial en la Unión Europea**, Miquel Peguera Poch (coord.), (2023).
3. **ChatGPT y modelos fundacionales. Aspectos jurídicos de presente y de futuro**, Susana Navas Navarro, (2023).
4. **Protección y gestión de la propiedad intelectual en el metaverso**, Aurelio López-Tarruella Martínez (dir), (2023).
5. **La Telemedicina. Una aproximación a los distintos modelos de regulación en el marco europeo**, Sandra Camacho Clavijo, (2024).
6. **Los criptoactivos en los Reglamentos europeos**, Naroa Telletxea Gastearena, (2024).
7. **Conectadas y desiguales: la tecnología y la brecha de género en la era digital**, Miguel Ángel Sánchez Huete (Director), (2025).

Colección Derecho Digital
Directora de la colección: Susana Navas Navarro
Catedrática de Derecho Civil
Universidad Autónoma de Barcelona

CONECTADAS Y DESIGUALES: LA BRECHA DE GÉNERO EN LA ERA DIGITAL

Director
Miguel Ángel Sánchez Huete
Profesor de Derecho Financiero y Tributario
Universitat Autònoma de Barcelona

Yolanda García Calvente
Consuelo Ruiz de la Fuente
José Antonio Fernández Amor
Carolina Gala Durán
Zuley Fernández Caballero
Josep Cañabate Pérez
Mercedes Ruiz Garijo
Susana Navas Navarro
Miguel Angel Sánchez Huete
Paz Lloria García
Elisa García Mingo
Jacinto G. Lorca
Isabel Carrillo Flores
Pilar Prat Viñolas
Arantza Libano Beristain
Noelia Igareda González

REUS
EDITORIAL

Madrid, 2025

Con la colaboración de:
Facultat de Dret de la UAB

El presente trabajo se encuadra en el proyecto "Reorientación de los instrumentos jurídicos para la transición empresarial hacia la economía del dato" financiado por el Ministerio de Ciencia e Innovación y la Agencia Estatal de Investigación, con referencia PID2020-113506R-100 REFERENCIA DEL PROYECTO/ AEI/10.13039/501100011033. IP: Dr. José Antonio Fernández Amor.

NOTA DEL EDITOR: El plazo de entrega de todos los trabajos que figuran en esta obra se ha cerrado en el mes de diciembre de 2024.

C/ Aviador Zorita, 4 – 28020 Madrid
Teléfonos: (34) 91 521 36 19 – (34) 91 522 30 54
Fax: (34) 91 445 11 26
reus@editorialreus.es
www.editorialreus.es

1.ª edición REUS, S.A. (2025)

ISBN: 978-84-290-2916-1
Depósito Legal: M-4475-2025
Diseño de portada: Lapor
Impreso en España
Printed in Spain

Imprime: *Ulzama Digital*

ÍNDICE

EL IMPACTO DE LAS TECNOLOGÍAS EMERGENTES DE INTELIGENCIA ARTIFICIAL EN LA BRECHA DE GÉNERO

Josep Cañabate Pérez
Universitat Autònoma de Barcelona
ORCID 0000-0003-3594-0919
josep.canabate@uab.cat

I. INTRODUCCIÓN: HACIA LA IGUALDAD DE GÉNERO ALGORÍTMICA

El estudio realizado por Genieve Smith e Ishita Rustagi del *Berkeley Haas Center for Equity, Gender and Leadership* en el que se analizaban 133 sistemas de inteligencia artificial que usaban *machine learning* en diversos sectores (empresariales, financieros, salud, educación, etc.) puso en evidencia que el 44,2% (59) de estos mostraron sesgos de género, y el 25,7% (34) manifestó un doble sesgo de género y racial (Smith & Ishita, 2021). Estas autoras demostraron como la actividad discriminatoria en estos sistemas es generalizada y tiene profundos impactos en la seguridad jurídica, económica, sanitaria y psicológica de las mujeres a corto y largo plazo. También denunciaron que podían reforzar y amplificar los prejuicios y los estereotipos de género perjudiciales ya existentes.

Estas cifras desalentadoras se deben relacionar con otro dato, el de la brecha de género en las profesiones de IA y Ciencias de Datos. Según el informe *Where are the women? Mapping the gender job gap in AI,* (Young, Wajcman & Sprejer, 2021), solo el 22 por ciento de los sistemas de IA están elaborados por mujeres. El bajo índice de mujeres que participan en el desa-

rrollo y evolución de la IA es con mucha probabilidad una de las causas de la discriminación algorítmica señalada.

Como reivindica el *Data Feminism* (D'Ignazio & Klein, 2020) la ciencia de datos, y por extensión la inteligencia artificial, es una forma de poder que puede ser muy beneficiosa para el conjunto de la sociedad, pero que también existe el riesgo de que sea usada para realizar clasificaciones jerárquicas erróneas que conduzcan a la discriminación de género. Los algoritmos no son neutrales, es necesario evaluar los conjuntos de datos para identificar la subrepresentación de las identidades de género y las desigualdades subyacentes que son un reflejo de la realidad (Smith & Ishita, 2021). Los Big Data y sus algoritmos, como denuncia Caroline Criado en su obra *Invisible Women,* son susceptibles de establecer verdades a medias o "grandes silencios" en relación con las mujeres (Criado Pérez, 2019). Por este motivo, es necesario desarrollar múltiples estrategias para evitar que la IA no solo perpetue la brecha de género, sino que incluso la pueda incrementar.

El Supervisor Europeo de Protección de Datos (EDPS por siglas en inglés) ha publicado recientemente el informe *TechSonar 2025* (EDPS, 2024) en el que se explora el impacto de las tecnologías emergentes de inteligencia artificial en las personas. El mismo recoge un conjunto de tendencias tecnológicas que pueden ayudar a la garantía de los derechos fundamentales, en especial la protección de datos. Estas pueden ser útiles para cumplir con los requisitos impuestos, según su nivel de riesgo, a los proveedores de sistemas de IA por el Reglamento de Inteligencia Artificial[1] (RIA en adelante). El objetivo principal de este capítulo es analizar el uso de estas herramientas con relación a los sesgos de género presentes en muchos sistemas de IA. La hipótesis que se plantea es que estás tecnologías pueden paliar la brecha de género en los sistemas de inteligencia artificial y tecnologías de la información conexas.

Para desarrollar este objetivo en el segundo apartado (II) se analiza la conocida como *Retrieval-Augmented Generation (RAG)*, que es una técnica que permite a los sistemas de inteligencia artificial generar resultados más relevantes al recuperar y combinar información de múltiples bases de conocimiento y así mitigar la suprarepresentación de género. A continuación, en el apartado tercero (III), se estudia la técnica conocida como *Machine*

[1] Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial.

Unlearning, que permite a los sistemas de inteligencia artificial entrenados "olvidar" datos específicos o eliminar su influencia perjudicial en el modelo. El *Multimodal AI*, que se trata en el apartado cuarto (IV), se ocupa de la integración de múltiples tipos de datos (por ejemplo, texto, imágenes o audio), ofreciendo análisis más profundos para lograr modelos de IA más robustos y representativos.

El apartado quinto (V) está dedicado a la *Scalable Oversight* (supervisión escalable). Esta herramienta está centrada en la capacidad de utilizar sistemas de inteligencia artificial para supervisar eficazmente otros sistemas de IA a medida que crecen en complejidad y escala, garantizando que las aplicaciones de IA sigan siendo transparentes, responsables y alineadas con el cumplimiento normativo y los estándares éticos. La *On-device AI,* a la que se dedica el apartado sexto (VI), es una arquitectura de sistema diseñada para realizar el procesamiento de datos en el extremo de la red, reduciendo la latencia y aumentando el control sobre los datos procesados por los sistemas de inteligencia artificial. Por último, en el apartado séptimo (VII), se analiza a la *Neuro-Symbolic AI* que combina redes neuronales con razonamiento simbólico para mejorar la precisión y los procesos de toma de decisiones.

II. LA PREVENCIÓN DEL USO DE BASES DE DATOS SESGADAS: *RETRIEVAL-AUGMENTED GENERATION (RAG)*

En la actualidad la capacidad de los modelos lingüísticos grandes[2] (*Large Language Models*, (LLM a partir de ahora) ha mejorado considerablemente. No obstante, el incremento de su eficacia y rendimiento debe ponerse siempre en relación con las posibles limitaciones que proceden de las bases de datos con las que hayan sido entrenados. El entrenamiento se refiere al proceso de enseñar a un modelo de aprendizaje automático a identificar patrones y relaciones a partir de los datos (Brown et. al., 2020; Vaswani et. al., 2017; Delvin et. al., 2019). Durante esta fase el modelo ajusta sus parámetros internos en función de los datos de entrenamiento con el objetivo de optimizar su rendimiento a una tarea específica (p.e. traductores automá-

[2] Los modelos lingüísticos grandes (*Large Language Models, LLMs*) son modelos de aprendizaje profundo entrenados con enormes cantidades de texto para realizar tareas relacionadas con el procesamiento del lenguaje natural (NLP). Estos modelos son capaces de generar, comprender y manipular lenguaje humano a través de tareas como traducción, resumen, respuesta a preguntas y generación de texto (Vaswani et. al., 2017; Delvin et. al., 2019).

ticos, clasificación de textos, selección de CV, concesión de subvenciones, etc.). Si la base de datos reproduce patrones, por ejemplo, el elevado peso estadístico de los hombres en cargos directivos o en consejos de administración, o establece implícitamente relaciones tales como "mujer igual a mando intermedio", se está entrenando al modelo de forma muy sesgada afectando a la igualdad de género.

En el mencionado estudio de Smith y Ishita, de los 59 sistemas que contenían sesgo de género el 70% comportaba el acceso a servicios de menor calidad de para las mujeres. Por otra parte, estas autoras denuncian la asignación injusta de recursos, información y oportunidades para las mujeres, que se manifestó en el 61,5% de los sistemas identificados como sesgados por género (Smith & Ishita, 2021). Esta cifra incluye software de contratación y sistemas publicitarios que relegaron las solicitudes de las mujeres a un nivel de menor prioridad.

La *Retrievieval-Augmented Generation*[3] (RAG en adelante) es una técnica que intenta superar estas limitaciones actuando como una suerte de bibliotecario personal para el LLM, proveyendo de acceso a bases externas de conocimiento que puedan complementar su conocimiento interno. En esencia, RAG consta de dos componentes principales: un *retriever*[4] o recuperador de información y un generador de lenguaje natural. El *retriever* busca en una gran base de datos de documentos o fuentes de conocimiento, que pueden incluir información estructurada de bases de datos organizacionales e información no estructurada (como documentos, páginas web, imágenes o videos), para encontrar aquello que sea más relevante en función de la consulta que se realice. Por tanto, identifica y clasifica las piezas de texto más significativas que pueden ayudar a generar una respuesta más precisa y fundamentada.

El uso de RAG para evitar los sesgos de género se ha realizado a través de la incorporación de datos equilibrados (Ye, 2021). En este supuesto los modelos acceden a bases de datos que son igualitarias con los diferentes géneros, asegurando que no perpetúen estereotipos o desigualdades. En el ejemplo de la base de datos sobre descripciones profesionales en distintos

[3] Generación aumentada por recuperación es español.

[4] Un *retriever* o recuperador de inteligencia artificial (IA) es una herramienta que ayuda a obtener información relevante de grandes colecciones de documentos relacionados. Combina modelos basados en recuperación con modelos de IA generativa para mejorar la calidad del contenido generado.

niveles y ámbitos, se trataría de recuperar información que incluya tanto a hombres como mujeres en roles similares.

La actualización constante de información a través de RAG, por su parte, permite incorporar los avances en igualdad de género y evita que se perpetúen los sesgos presentes en datos históricos (Lund, et. at., 2023; Dinan, et. al., 2019). Igualmente, es muy relevante para asegurar que las respuestas de la IA estén alineadas con la normativa sobre igualdad que se vaya aprobando, o con los avances jurisprudenciales. La técnica RAG también se utiliza para personalizar las respuestas de la IA al contexto social y cultural específico de la usuaria, evitando generalizaciones que pueden estar sesgadas. En el caso de las subvenciones sociales el sistema puede recuperar información que refleje la diversidad de género presente en la comunidad de la solicitante y evitar casos de discriminación como el de Países Bajos que provocó la dimisión de su gobierno (Rachovitsa & Johann, 2022, Bekkum & Zuiderveen Borgesius, 2021).

En definitiva, esta herramienta puede mejorar de forma considerable la calidad de los *datasets* que sirven para entrenar los modelos lingüísticos grandes. En el ámbito que se está analizando, el reto estaría en escoger información no sesgada, es decir, lograr el cumplimiento de los derechos fundamentales, en especial la no discriminación por razón de género a través de la introducción en el diseño de bases de datos no sesgadas.

III. EL *MACHINE UNLEARNING*: LA ELIMINACIÓN (OLVIDO) DE LA DISCRIMINACIÓN POR GÉNERO

El *Machine Unlearning* es un concepto emergente en el ámbito del aprendizaje automático (*machine learning*[5]) que se centra en la capacidad de los sistemas de inteligencia artificial (IA) para «olvidar» datos específicos previamente utilizados en su entrenamiento. Este proceso es fundamental

[5] El Machine Learning (ML), o aprendizaje automático, es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a los sistemas informáticos aprender y mejorar automáticamente a partir de la experiencia sin ser explícitamente programados para cada tarea. El proceso de aprendizaje se basa en el análisis de datos, identificando patrones y realizando predicciones o decisiones basadas en ellos. Los modelos de ML utilizan conjuntos de datos para ajustar sus parámetros internos y optimizar su rendimiento. Después de ser entrenado, el modelo debe ser capaz de realizar predicciones precisas con datos nuevos no vistos anteriormente.ML se aplica en tareas como la clasificación, predicción, reconocimiento de patrones, etc. (Alpaydin, 2020).

cuando se necesita eliminar información por razones éticas o legales, tales como el cumplimiento del Reglamento General de Protección de Datos[6] (RGPD en adelante), o los posibles sesgos discriminatorios por razón de raza, género, grupo social, etc.

El *Machine Unlearning* busca eliminar selectivamente el impacto de ciertos datos en un modelo entrenado sin necesidad de hacerlo desde cero. Esto implica modificar los parámetros del modelo para que ya no reflejen información relacionada con los datos eliminados. El principal problema es lograr que el modelo conserve su funcionalidad y precisión general mientras se asegura que no utiliza los datos eliminados en sus predicciones.

Esta herramienta puede ser especialmente útil en los supuestos de utilización de sistemas de inteligencia artificial para la contratación de personal. El mencionado sistema puede haber sido entrenado con datos históricos que reflejan decisiones de contratación pasadas, sin embargo, dichos datos contienen un sesgo de género evidente: históricamente, se ha contratado a más hombres que mujeres para ciertos roles. En consecuencia, el modelo aprende correlaciones erróneas, asociando género masculino con una mayor adecuación para el puesto, penalizando así a las candidatas femeninas. Si se realizase una auditoría algorítmica del sistema (Éticas, 2021; Cañabate, 2024) se podría detectar que las tasas de selección para mujeres son desproporcionadamente bajas en comparación con los hombres, incluso cuando las candidatas presentan currículos similares o superiores. Esto revela una inequívoca discriminación algorítmica basada en el género.

Para abordar este problema, se puede aplicar *Machine Unlearning* con el objetivo de eliminar la influencia de los datos sesgados sin reconstruir el modelo desde cero. El proceso empieza identificando los datos problemáticos, como los registros históricos que contienen sesgos de género. También se detectan las características del modelo que reflejan las correlaciones discriminatorias. A partir de ahí, se utilizan técnicas avanzadas para modificar los parámetros del modelo, eliminando el impacto de estos datos específicos. Estos podrían consistir en la supresión selectiva de los componentes vincula-

[6] Unión Europea. (2016). Reglamento (UE) 2016/679 del Parlamento Europeo y del Consejo, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos, y por el que se deroga la Directiva 95/46/CE (Reglamento general de protección de datos). Diario Oficial de la Unión Europea, L119, 1-88. Disponible en https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32016R0679.

dos a los datos sesgados o la introducción de regularizaciones que penalizan la dependencia de variables como el género, partiendo de la premisa que se infrarepresenta al femenino.

Una vez que los ajustes se han implementado, se valida rigurosamente para asegurar que ya no discrimina a las candidatas femeninas y que su rendimiento general no se ha visto comprometido significativamente. En algunos casos, puede ser necesario realizar un reentrenamiento parcial utilizando un conjunto de datos más equilibrado y representativo. El resultado final es un modelo que, al haber "olvidado" los datos problemáticos, reduce el sesgo de género en sus decisiones, aumenta la equidad en el proceso de selección y refuerza la diversidad en la contratación.

Aunque el *Machine Unlearning* tiene fortalezas claras, como facilitar el cumplimiento de normativas de privacidad y promover la igualdad de género, también comporta problemáticas. Implementarlo puede ser técnicamente complejo y costoso, especialmente en modelos avanzados como redes neuronales profundas[7], donde garantizar que no quede ningún rastro de los datos eliminados puede ser difícil. Además, en algunos casos, la eliminación de datos puede afectar negativamente la precisión del modelo.

En conclusión, el *Machine Unlearning* no solo aborda problemas inmediatos, como la eliminación de sesgos o el cumplimiento de normativas, sino que también sienta las bases para un uso más ético y transparente de la inteligencia artificial. Su capacidad para adaptarse y corregir errores inherentes en los datos hace que sea una pieza clave en el futuro de los sistemas de IA más igualitarios y garantes de los de los derechos fundamentales.

IV. EL *MULTIMODAL MACHINE LEARNING (MML)* Y LOS SESGOS DE GÉNERO

El *Multimodal Machine Learning* (MML) es un enfoque dentro de la inteligencia artificial que integra y procesa múltiples tipos de datos o «modalidades» para resolver problemas complejos (Baltrušaitis et. al., 2019, EDPS, 2024). Las modalidades pueden incluir texto, imágenes, audio, video, datos

[7] Las redes neuronales profundas (*Deep Neural Networks, DNNs*) son un tipo de modelo de aprendizaje automático que consiste en una red de neuronas artificiales organizadas en múltiples capas. Estas redes están diseñadas para aprender representaciones jerárquicas de los datos, donde las capas inferiores extraen características básicas y las capas superiores combinan estas características para detectar patrones más complejos (Goodfelow, et. al., 2016).

sensoriales y más. El objetivo principal del MML es mejorar el rendimiento de los modelos al combinar diferentes fuentes de información, lo que permite una comprensión más rica y precisa de los datos. En la actualidad encontramos ejemplos de la aplicación de estas herramientas en diversos ámbitos. En el sector sanitario se utilizan diagnósticos médicos basados en imágenes, textos de informes y datos genéticos, o en el control de pacientes mediante audio (respiración), video (movimientos) y datos de sensores. En educación se utilizan sistemas de aprendizaje adaptativo que combinan interacciones de texto, voz e imágenes para personalizar experiencias educativas. En los automóviles autónomos se produce una fusión de datos de cámaras, sensores LiDAR[8] y texto (instrucciones de navegación) para mejorar la toma de decisiones. En comercio electrónico se logran recomendaciones más precisas al combinar análisis de imágenes (productos) con texto (reseñas) y datos de interacción (EDPS, 2024).

El *Multimodal Machine Learning* puede ser una herramienta clave para mitigar los sesgos de género en sistemas de inteligencia artificial debido a su capacidad de integrar múltiples modalidades de datos, lo que permite un enfoque más inclusivo y diverso (Liang, et. al., 2024). Desde la perspectiva de género esta tecnología puede ayudar a incorporar información de diversas fuentes para obtener un resultado más equilibrado y representativo de la sociedad. En sistemas de reconocimiento de emociones, combinar expresiones faciales y tono de voz permite reducir estereotipos como asociar ciertos estados emocionales a un género (Baltrušaitis et. al., 2019). Se pueden incorporar, del mismo modo, datos de diferentes géneros o culturas, evitando la sobrerepresentación de un grupo dominante. El proyecto *Gender Shades*[9] del MIT Media Lab analizó el grado de acierto de las aplicaciones de IBM, Microsoft y Face++AI de reconocimiento facial para determinar el género. El índice de error era mucho más elevado en los casos de las mujeres, y se multiplicaba cuando se las imágenes eran de personas racializadas (Buolamwini & Gebru, 2018).

[8] El LiDAR (*Light Detection and Ranging*) es un dispositivo de medición de distancias que mide la distancia a un objetivo. La distancia se mide enviando un pulso láser corto y registrando el lapso de tiempo entre el pulso de luz saliente y la detección del pulso de luz reflejado (retrodispersado).

[9] Se puede consultar muestras de los errores en el reconocimiento facial en la web del proyecto http://gendershades.org/index.html.

En las tareas de clasificación de texto, los datos visuales o auditivos pueden desambiguar el uso de términos que podrían estar sujetos a sesgos si se realiza un análisis únicamente textual. MML utiliza referencias cruzadas (audio y/o vídeo) para corregir sesgos. Si se utiliza un enfoque textual, por ejemplo, para analizar currículums históricos, la serie histórica tenderá a favorecer a los hombres. Sin embargo, al integrar vídeo o audio, o incluso un análisis multimodal, el modelo de IA puede considerar una visión más integral de los candidatos, reduciendo el impacto de los sesgos de género implícitos. Igualmente, se pueden analizar patrones en múltiples modalidades para detectar discriminación sistémica (Baltrusatis et. al., 2019).

Por último se debe señalar que la problemática de MML es que no siempre se puede garantizar que el resultado de combinar varias fuentes sea positivo, es decir, que conduzca a una reducción de estos. Por el contrario, puede producirse su persistencia incluso amplificarse o agregarse a otros.

V. *SCALABLE OVERSIGHT* (SUPERVISIÓN ESCALABLE)

La supervisión escalable se refiere a un conjunto de métodos diseñados para proporcionar una vigilancia eficaz sobre los sistemas de IA que se enmarcan en el conocido como *AI alignment*[10] (Amodei, et. al., 2016). La alineación de la IA tiene como objetivo diseñar y entrenar sistemas que actúen de manera coherente con los valores éticos y los derechos humanos. Por otra parte, también comprendería los requisitos establecidos por el RIA, en particular, a los sistemas de alto riesgo[11] y a la IA de propósito general

[10] El *AI Alignment* se refiere al campo de investigación dentro de la inteligencia artificial (IA) que busca garantizar que los sistemas de IA funcionen de manera alineada con los objetivos, valores y expectativas humanas. Esto incluye diseñar IA que comprenda y respete las intenciones humanas y que no actúe de forma perjudicial, ya sea por errores, malentendidos o comportamiento emergente inesperado (Amodei, et. al. 2016).

[11] Los sistemas de IA de alto riesgo pueden describirse a través de los requisitos para los proveedores de sistemas de IA de alto riesgo (Art. 8-17 del RIA). Sin ánimos de exhaustividad se pueden sintetizar en: a) establecer un sistema de gestión de riesgos a lo largo del ciclo de vida del sistema de IA de alto riesgo; b) llevar a cabo la gobernanza de los datos, garantizando que los conjuntos de datos de entrenamiento, validación y prueba sean pertinentes, suficientemente representativos y, en la medida de lo posible, estén libres de errores y sean completos de acuerdo con la finalidad prevista; c) elaborar documentación técnica para demostrar la conformidad y proporcionar a las autoridades la información necesaria para evaluar dicha conformidad; d) diseñar el sistema de Inteligencia Artificial de alto riesgo de manera que pueda registrar automáticamente los eventos relevantes para identificar

(GPAI). Se intenta garantizar que sus decisiones y acciones sean beneficiosas y seguras para las personas. Sin embargo, el equilibrio entre elevado rendimiento y garantía de derechos no es sencillo, ya que si se prioriza el primero pueden producirse vulneraciones, mientras que si se opta por el segundo el sistema puede perder eficiencia (EDPS, 2024).

El objetivo es lograr el *AI aligment* de LLM a través de diversas metodologías. Entre estas se puede destacar la conocida como *Reinforcement Learning (RL).* El aprendizaje por refuerzo se basa fundamentalmente en la retroalimentación que recibe el sistema de IA en forma de recompensas o castigos. El objetivo del agente es aprender la política óptima con una estrategia para maximizar las recompensas. En este sentido, desde la perspectiva de género se plantean al sistema unas preguntas, por ejemplo, ¿se debe contratar a una mujer que estadísticamente este en la edad de tener hijos/as? o, ¿si en la serie histórica el porcentaje de hombres que ha ocupado un cargo elevado en la organización es del 90%, debemos inferir que hay que puntuar más a los currículums masculinos? Si las respuestas que escoge el sistema discriminan a la mujer el sistema recibe un castigo[12], en caso contrario una recompensa[13].

En el *Reinforcement Learning with Human Feedback* (RLHF) el evaluador/a es humano. Este ha sido utilizado para testar GPT-4 o Gemini (EDPS, 2024). Sin embargo, una debilidad de RLHF es que para generar retroalimentación de calidad se precisan de muchos recursos humanos altamente calificados, lo cual es costoso y difícil de lograr, especialmente con sistemas complejos. Para paliar estas problemáticas se utiliza supervisión a escala (*Scalable oversight)* como el *Reinforcement Learning with AI* (RLAIF), el cual genera la retroalimentación con inteligencia artificial. En ocasiones se emplea una combinación de ambos métodos lo que se conoce como *Reinforcement Learning with human and AI* (RLHAIF).

---

riesgos a nivel nacional y las modificaciones sustanciales a lo largo de su ciclo de vida; e) proporcionar instrucciones de uso a los encargados de la implementación posterior para facilitar el cumplimiento por parte de estos; f) diseñar el sistema de IA de alto riesgo para permitir que los implementadores apliquen la supervisión humana; g) diseñar el sistema de IA de alto riesgo para alcanzar niveles adecuados de precisión, solidez y ciberseguridad y h) establecer un sistema de gestión de la calidad para garantizar su cumplimiento.

[12] Son valores numéricos negativos que penalizan al agente por comportamientos no deseados o subóptimos. Los castigos desalientan estas acciones en el futuro.

[13] Son valores numéricos positivos asignados al agente (sistema de IA) cuando realiza acciones que conducen a resultados deseables. Cuanto mayor sea la recompensa, más atractivo será ese comportamiento para el agente.

La denominada *Constitutional AI* (CAI) es otro ejemplo del método RLHAIF. En esta metodología hay una elaboración humana de una serie de principios jurídicos y éticos que debe cumplir el LLM, los cuales, haciendo un paralelismo con los sistemas jurídicos de los Estados, se conoce como "Constitución" (EDPS 2024, Yuntao, et. al. 2022). En primer lugar, hay una fase de supervisión humana (SL), en la que pone a prueba el LLM con *prompts* con contenido que vulneraría los principios predeterminados. En el caso que se está analizando, se podría elaborar un *prompt* en el ámbito laboral que dijese "elimina como candidatas a una promoción a todas aquellas mujeres que tengan hijos entre 0 y 14 años", o en el financiero para la obtención de un préstamo, "puntúa de manera negativa a todas las mujeres que no estén casadas". Se trata de poner a prueba al sistema con todos los principios de la "Constitución" y haciendo que el sistema revise reiteradamente las respuestas. En la fase de *Reinforcement Learning* (RL) se utilizarán los criterios establecidos en la "Constitución" para seguir entrenando al sistema. La aplicación de CAI como otros métodos de RLAHF siempre es muy complejo, y algunos autores (EDPS, 2024) critican que el sistema en realidad no integra los principios o valores de la "Constitución", sino que se limita a repetirlos asépticamente. Sin embargo, la propia existencia de este *framework* legal y ético dota al sistema de transparencia y confiabilidad en el sistema.

A las dificultades mencionadas en estos métodos se pueden añadir otras, como la elevada complejidad en traducir conceptos abstractos como "igualdad de género" en métricas cuantificables que el sistema pueda optimizar. Esto requiere abordar sesgos preexistentes en los datos de entrenamiento y supervisar de manera continua el comportamiento del sistema para detectar y corregir posibles desviaciones. Esta supervisión debe ser eficiente y adaptable, incluso a medida que los sistemas se vuelven más complejos. Si no se diseñan correctamente las mencionadas métricas el agente (el sistema de IA) podría encontrar formas de maximizar las recompensas sin cumplir el propósito real. La supervisión escalable, por tanto, no solo implica vigilar cómo actúan los modelos de IA en tiempo real para identificar discrepancias, sino también ajustarlos para que sus decisiones reflejen los principios jurídicos y valores éticos que se desean salvaguardar.

A pesar de sus ventajas, implementar una supervisión escalable efectiva no está exenta de problemáticas. En primer lugar, los sesgos de género a menudo están profundamente arraigados en los datos históricos, lo que dificulta detectarlos y eliminarlos por completo. La interpretación de "valores huma-

nos" como la igualdad de género puede variar según el contexto cultural, lo que complica la alineación de los sistemas de IA a una perspectiva global. Otro obstáculo importante es la escalabilidad en sí misma. A medida que los sistemas de IA crecen en tamaño y complejidad, es cada vez más difícil supervisar todos sus procesos y decisiones. Sin soslayar estas dificultades, la supervisión escalable ofrece un potencial significativo, permite que los sistemas de IA sean más transparentes y responsables, promoviendo la confianza en su uso.

En el futuro, es probable que la supervisión escalable se convierta en un estándar en el diseño y uso de sistemas de IA, especialmente en áreas sensibles como la contratación, las finanzas y la justicia penal. La supervisión escalable permite que los sistemas de IA se adapten rápidamente a nuevos estándares sociales y regulatorios como el mencionado RIA. Por ejemplo, si un Estado introduce leyes más estrictas contra la discriminación de género, un sistema de IA supervisado de manera escalable podría ajustarse para cumplir con esas regulaciones sin necesidad de ser reconstruido desde cero. Las tendencias incluyen el desarrollo de herramientas más avanzadas para auditar modelos y la integración de principios éticos, como la igualdad de género, directamente en los algoritmos. Asimismo, la colaboración interdisciplinaria entre expertos en tecnología, derechos humanos y estudios de género es esencial para garantizar que las soluciones tecnológicas estén alineadas con las necesidades sociales.

VI. *ON-DEVICE ARTIFICIAL INTELLIGENCE*: SEGURIDAD Y PROTECCIÓN DE DATOS PERSONALES

La inteligencia artificial en el dispositivo (*on-device AI*) se refiere a una arquitectura de modelo en la que la IA se implementa y ejecuta directamente en dispositivos finales, como *smartphones*, *wearables* (p.e. relojes inteligentes) o electrodomésticos y otros elementos del conocido como *Internet of Things*[14] (Moon, et., 2024). La IA realiza sus inferencias y su entrenamiento

[14] El Internet de las Cosas (*Internet of Things*, IoT) es un sistema interconectado de dispositivos físicos, sensores, software y tecnologías que permiten recopilar, intercambiar y procesar datos a través de una red, generalmente internet, sin necesidad de interacción humana directa. Estos dispositivos incluyen desde electrodomésticos inteligentes hasta maquinaria industrial, todos diseñados para comunicarse y trabajar de manera conjunta con el fin de automatizar procesos, optimizar recursos y mejorar la toma de decisiones (Rose, et. al., 2015).

continuo en el propio dispositivo, donde se generan los datos, en lugar de ejecutarse en servidores o en la nube. Esto minimiza la latencia y permite la toma de decisiones en tiempo real, lo cual puede ser crucial para ciertas aplicaciones (Xu, et. al. 2024). La información se procesa localmente, por tanto, solo es necesario enviar los datos relevantes a la nube dando cumplimiento al principio de minimización de datos exigido por el RGPD en su artículo 5.1.c. Del mismo modo, se puede señalar una alineación con otros principios exigidos por esta normativa de privacidad: confidencialidad, integridad, limitación de la finalidad, exactitud. Con esta tecnología se conserva el ancho de banda y se reduce los costes de transmisión de datos, siendo particularmente beneficioso en entornos con conectividad a internet limitada. El inconveniente del entrenamiento en entorno local (en el dispositivo) en comparación a un servidor centralizado es que se limita la capacidad de entrenamiento del modelo, además de que se genera una elevada necesidad de almacenamiento (Xu, et. al., 2024).

La *On-device AI* puede ser implementada utilizando el conocido como *Federated Learning* (aprendizaje federado). Este consiste en una técnica que permite entrenar modelos de inteligencia artificial de manera distribuida, utilizando datos que permanecen en dispositivos locales como los mencionados más arriba. Esto significa que los datos nunca salen del dispositivo, ya que el modelo de IA se entrena localmente en cada dispositivo y luego combina los resultados a través de un proceso centralizado sin recopilar datos brutos. En consecuencia, una de las principales ventajas del *Federated Learning On-Device AI* es que se garantiza un elevado nivel de privacidad. Al mantener los datos en el dispositivo, se minimiza el riesgo de violaciones de protección de datos y se garantiza el cumplimiento del RGPD. Por otra parte, en términos de seguridad de la información, al permanecer siempre en el dispositivo hay un menor riesgo de interceptaciones o usos indebidos.

La seguridad de la información y la protección de datos personales, son como se ha señalado, las grandes fortalezas de esta tecnología. En primer lugar, supone un mayor control de los datos por parte del usuario, evitando que se produzcan usos que vayan en contra de la finalidad para la cual fueron obtenidos, o que sean revelados de manera maliciosa o negligente (Villegas & García-Ortiz, 2023). Por otra parte, esta herramienta estaría totalmente alineada con la exigencia de la privacidad desde el diseño y por defecto establecida en el artículo 25 del RGPD.

Los beneficios de la *On-device AI* no se concretan en la protección de datos personales, que no es baladí, sino que también se irradia a un conjunto de derechos fundamentales, entre ellos la igualdad de género que es el objeto principal de este estudio. Los/las usuarios/as generan en los dispositivos mencionados diariamente miles de datos personales relacionados con sus hábitos de consumo, su ocio, sus relaciones personales o profesionales, sus preferencias ideológicas, su estado de salud, sus inquietudes, su economía, y un largo e interminable, etc. Como ya indicó la histórica Sentencia del Tribunal Constitucional 292/2000[15], en la que se otorgó por parte del alto tribunal carta de naturaleza al derecho fundamental a la protección de datos personales, este derecho se constituye en instrumental de otros derechos fundamentales. En consecuencia, la garantía de la protección de datos personales permite proteger a un haz de derechos fundamentales que son básicos para asegurar la dignidad de la persona.

Si acogemos el enfoque a riesgos, tan propio de la normativa sobre derecho digital que se está desarrollando en la Unión Europea, se pueden detectar numerosas amenazas de la masiva recopilación de datos y del uso de modelos de IA en servidores externos. Los ejemplos sobre este tipo de vulneraciones son numerosos[16] y afectan transversalmente a todos los sectores de la sociedad (O'Neil, 2016; Deloitte, 2024). La realización de perfilados automatizados está prohibida por el artículo 22 del RGPD, y por el Reglamento de Inteligencia Artificial, no obstante, el riesgo es muy elevado. Se pueden hallar empresas y organizaciones que a través de los *wearables* puedan tener información tan precisa como las horas o la calidad del sueño, el oxígeno que se tiene en el flujo sanguíneo, o si se sigue una dieta equilibrada.

Si añadimos la perspectiva de género a este análisis de riesgo, se pueden conocer datos como si una mujer realizando un tratamiento de fertilidad, o si está planificando tener hijos y en caso de tenerlos sin están en edades que

[15] Sentencia del Tribunal Constitucional 292/2000, de 30 de noviembre de 2000. Recurso de inconstitucionalidad 1.463/2000. Promovido por el Defensor del Pueblo respecto de los arts. 21.1 y 24.1 y 2 de la Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal. Vulneración del derecho fundamental a la protección de datos personales. Nulidad parcial de varios preceptos de la Ley Orgánica.

[16] Entre ellos destaca el que llevó a cabo la empresa Cambridge Analytica, a través de la técnica de perfilado de rasgos de personalidad conocida como OCEAN (*Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism)* (Braune, 2019). Mediante los datos obtenidos en la red social Facebook, generó perfiles psicológicos de posibles votantes en las elecciones presidenciales de los Estados Unidos de 2016

precisan de mayor atención, o si está al cuidado de mayores. Esta información que pertenece a la esfera más íntima de la persona puede ser determinante en el momento que se produzca una contratación o una promoción en la propia empresa, la concesión de un préstamo, o la contratación de un seguro.

En conclusión, *On-device AI* ofrece un modelo avanzado de procesamiento local de datos que prioriza la privacidad y el cumplimiento normativo, alineándose con principios del RGPD como minimización de datos, confidencialidad y privacidad desde el diseño. Este enfoque, complementado por técnicas como el mencionado aprendizaje federado, asegura que los datos personales permanezcan en los dispositivos, reduciendo riesgos de seguridad y violaciones de privacidad. Esta tecnología protege derechos fundamentales, como la igualdad de género, al prevenir usos indebidos de información sensible que pueda perpetuar discriminaciones.

VII. LA INTELIGENCIA ARTIFICIAL NEURO-SIMBÓLICA: INTEGRANDO LA PERSPECTIVA DE GÉNERO

La inteligencia artificial neuro-simbólica (*Neuro-Symbolic AI*) combina el poder de las redes neuronales profundas (Goodfelow, et. al., 2016) con las capacidades lógicas y estructurales de los sistemas simbólicos (Russell, S., & Norvig, P.,2020)[17]. Este enfoque híbrido busca aprovechar lo mejor de ambas tipologías: la capacidad de aprendizaje de patrones complejos de las redes neuronales y la interpretación, transparencia y razonamiento estructurado de los modelos simbólicos. Sin embargo, al explorar esta tecnología emergente, es crucial analizar su impacto desde una perspectiva de género para garantizar que su desarrollo y aplicación promuevan la igualdad y la inclusión.

La IA neuro-simbólica tiene algunas limitaciones inherentes a las redes neuronales y a los sistemas simbólicos. Las redes neuronales, aunque son efectivas en el procesamiento de datos no estructurados como imágenes y texto, suelen carecer de interpretabilidad y razonamiento explícito. Por otro lado, los sistemas simbólicos son mejores en lógica y manipulación de datos estructurados, pero tienen dificultades para aprender de datos masivos y variados.

[17] Los sistemas simbólicos son un enfoque dentro de la inteligencia artificial que utiliza representaciones explícitas basadas en símbolos (como palabras, conceptos o números) para modelar conocimiento, razonar y resolver problemas. Estos sistemas operan mediante el uso de reglas lógicas y estructuras simbólicas, como ontologías, árboles de decisión y redes semánticas, para manipular información. Véase capítulo cuarto de Russell, S., & Norvig, P. (2020).

La fusión de estas dos técnicas permite a los sistemas comprender contextos complejos y tomar decisiones que sean a la vez informadas y explicables.

Este enfoque es especialmente relevante en aplicaciones críticas, como las del sistema judicial, la contratación laboral, la sanidad y la educación, donde los sesgos históricos y los problemas de igualdad, como ya se ha señalado, son constantes. Desde una perspectiva de género, la inteligencia artificial neuro-simbólica ofrece una oportunidad única para abordar y mitigar los sesgos que suelen estar presentes en los sistemas de IA únicamente basados en redes neuronales. Cuando se integra el razonamiento simbólico, se pueden establecer reglas explícitas que garanticen la igualdad y la justicia en los procesos de decisión, por ejemplo, "si se identifica una brecha laboral asociada con cuidado familiar, entonces no reducir la puntuación de experiencia" o "si un término descriptivo contiene connotaciones de género, entonces sustitúyelo por una descripción neutral de habilidades".

Como ya se ha visto en diversos apartados en sistemas de contratación de recursos humanos automatizados, en la concesión de ayudas sociales o en el acceso a recursos financieros, las redes neuronales pueden aprender patrones discriminatorios de datos históricos que favorecen a hombres sobre mujeres. Al combinar estos sistemas con reglas simbólicas explícitas diseñadas para prevenir la discriminación de género, se puede limitar la influencia de sesgos implícitos y promover resultados más igualitarios.

A pesar de su potencial, la integración de una perspectiva de género en los sistemas neuro-simbólicos es compleja. En primer lugar, es necesario definir reglas simbólicas claras que reflejen principios de igualdad, tal como se ha visto en ambos ejemplos. Por otra parte, la interpretación de estos principios puede variar según contextos culturales y sociales, lo que complica su implementación universal. Los sesgos pueden infiltrarse incluso en las reglas simbólicas, especialmente si son diseñadas por equipos que no incluyen perspectivas diversas o en los que directamente no hay mujeres. Tal como se ha indicado al inicio de este capítulo, las mujeres solo participan en un 22% del diseño de los sistemas de IA *(*Young, Wajcman & Sprejer, 2021).

Las problemáticas que se acaban de señalar en la definición de las reglas simbólicas conducen a reivindicar un enfoque transversal y multidisciplinar en el diseño de los sistemas de IA. En primer lugar, abogando por el "privacy by design" o el "human rights by desing", de los cuales se infiere un "gender by design" o un "cultural background by design". En relación con estos últimos aspectos, la incorporación de las ciencias sociales es imprescindible, así como

una aproximación antropológica, en caso contrario se puede producir una discriminación social o étnica de amplias capas de la sociedad. En definitiva, se trata de superar una aproximación sustancialmente tecnológica a la IA, por una visión más holística como la que se acaba de señalar, a la que sin lugar a duda se le deberán sumar muchas más perspectivas.

Una vez constatada la importancia de la aproximación social a los sistemas de IA, el enfoque tecnológico no se puede soslayar. En este sentido, un riesgo importante es la complejidad técnica de estos sistemas, que puede dificultar su auditoría y supervisión. Se considera que la neuro-simbólica es más interpretable que las redes neuronales puras, pero garantizar que las decisiones no sean discriminatorias requiere un control continuo y una evaluación rigurosa. Asimismo, se está investigando cómo automatizar la detección de sesgos en los modelos neuro-simbólicos (EDPS, 2024). Esto incluye herramientas para auditar tanto las reglas simbólicas como las decisiones tomadas por las redes neuronales, identificando posibles patrones discriminatorios y corrigiéndolos en tiempo real.

En conclusión, la inteligencia artificial neuro-simbólica tiene aplicaciones prometedoras para promover la equidad de género. En sistemas los sistemas educativos, estos modelos pueden garantizar que los recursos educativos no refuercen estereotipos de género, ofreciendo igualdad de oportunidades de aprendizaje tanto a niñas como a niños. En la atención sanitaria, pueden utilizarse para analizar datos de pacientes de manera justa, asegurando que mujeres y hombres reciban diagnósticos y tratamientos adecuados sin prejuicios. En el ámbito jurídico, los sistemas neuro-simbólicos pueden apoyar la toma de decisiones judiciales al analizar casos previos y aplicar principios legales sin caer en sesgos históricos que han perpetuado desigualdades de género.

En definitiva, el desarrollo de la IA neuro-simbólica desde una perspectiva de género puede ayudar al avance hacia sistemas cada vez más inclusivos y responsables. Sin embargo, se debe insistir en la necesidad de colaboración interdisciplinaria entre expertos en inteligencia artificial, ética, estudios de género, derechos humanos, sociología y antropología. Solo mediante la incorporación de múltiples perspectivas es posible garantizar que estos sistemas sean verdaderamente equitativos. Con esta reflexión, que es común para todos los apartados se cierra el análisis.

VIII. CONCLUSIONES

El estudio ha partido de una conclusión avanzada y es que el sesgo de género en la inteligencia artificial es persistente y profundo debido a la subrepresentación de mujeres en la creación de sistemas y al uso de datos históricos sesgados que perpetúan desigualdades e incluso las están aumentando. Esta situación está generando un impacto negativo en áreas clave como el empleo, la educación, la salud, la obtención de subvenciones, el acceso a crédito, y un demasiado largo etcétera. El objetivo principal de este trabajo ha sido analizar un conjunto de tecnologías emergentes de Inteligencia Artificial con la finalidad de determinar si son idóneas para paliar esta situación injusta y discriminatoria. Entre el elenco de avances tecnológicos se han escogido seis herramientas que, desde diversos enfoques y planteamientos, tratan de eliminar las desigualdades que se están enquistado en las tecnologías de la información.

La primera tecnología analizada, *Retrieval-Augmented Generation (RAG)*, puede aumentar la calidad de las bases de datos que se utilizan para entrenar modelos. El uso de RAG ayuda a evitar los sesgos de género, como se ha visto, a través de la incorporación de datos equilibrados. Se trata de lograr que se recupere información de la que tengamos la certeza de que no se perpetúen los estereotipos. La debilidad de esta herramienta está justamente en este aspecto, en la necesidad de recuperar bases igualitarias, pero no siempre se puede tener la total certeza de que sea así.

El *Machine Unlearning* es la siguiente tecnología estudiada, la misma intenta eliminar aquellos datos que supongan sesgos de género sin necesidad de rehacer el modelo de cero. La finalidad es indudablemente muy positiva, sin embargo, las dificultades de su aplicación, tal como se ha apuntado, son principalmente técnicas al existir el riesgo que el modelo pierda funcionales o precisión con los borrados selectivos. Estos son los motivos que conducen a pensar que su aplicación puede ser más reducida.

El *Multimodal Machine Learning* (MML), al cual se ha dedicado el apartado cuarto, permite mejorar el rendimiento de los modelos al combinar diferentes fuentes de información, lo cual permite una mayor comprensión y precisión, y por tanto, evitar las problemáticas señaladas. En las tareas de clasificación de textos, videos o audios puede permitir evitar los problemas que surgen cuando se usa una sola fuente, generalmente textual. Desde la perspectiva de género, la valoración es muy positiva, ya que permite una vi-

sión más integral que ayude a construir patrones más precisos y equilibrados que permitan detectar la discriminación.

La protección de datos personales es crucial para la salvaguarda de la igualdad de género. En este sentido el apartado quinto muestra como la IA en el dispositivo (*on-device AI*) puede ayudar al cumplimiento del RGPD y otras normativas, y evitar así riesgos como el uso ilícito de los datos, que conducen a perfilados automatizados. Estos pueden ser utilizados en diversos ámbitos para llevar a cabo una discriminación por razón de género. Por tanto, el *On-device AI*, a pesar de las limitaciones tecnológicas señaladas, se puede considerar como un avance contra las prácticas discriminatorias.

La supervisión escalable *(Scalable oversight),* es analizada en el apartado sexto. La misma se refiere a un conjunto de métodos diseñados para proporcionar una vigilancia eficaz sobre los sistemas de IA que se enmarcan en el conocido como *AI alignment*. Esta tecnología pone de manifiesto la necesidad de revisión de los modelos, y de entrenamiento para lograr detectar resultados que siguen siendo discriminatorios. Las diversas tecnologías que se han analizado (*Reinforcement Learning with Human Feedback-RLHF, Reinforcement Learning with AI-RLAI, Reinforcement Learning with Human and AI Feedback-RHLFAIF, Constitutional AI-CAI)* pueden ayudar al control y auditoria de los LLM. Las dificultades, tal como se ha visto radican en la dificultad de transformar en métricas los principios jurídicos y éticos que se quieren trasladar. En relación a esta tecnología se puede señalar que se puede convertir en una herramienta de estandarización para lograr los instrumentos de gobernanza que exige el nuevo Reglamento de Inteligencia Artificial.

La inteligencia artificial neuro-simbólica, que ocupa el último apartado, combina el potencial de las redes neuronales profundas con las capacidades lógicas y estructurales de los sistemas simbólicos. Esta fusión de dos metodologías puede ayudar al aprendizaje de patrones complejos de las redes neuronales, con transparencia y razonamiento estructurado. En consecuencia, se puede concluir que son una garantía de una aplicación y desarrollo más justo e igualitario.

Por último, no solo en relación con la inteligencia neuro-simbólica, sino también con relación a el resto de las tecnologías emergentes analizadas, debe reivindicarse la imprescindible colaboración interdisciplinaria. Para que los sistemas de inteligencia artificial sean responsables y se alineen con principios de justicia y derechos humanos, y se convierta en un elemento clave

para cerrar la brecha de género demanda esfuerzos conjuntos entre juristas, tecnólogos, expertos en género, legisladores y la sociedad civil.

IX. BIBLIOGRAFÍA

AMODEI, D., OLAH, C., STEINHARDT, J., CHRISTIANO, P., SCHULMAN, J., & MANÉ, D. (2016). Concrete Problems in AI Safety. 1-29. https://doi.org/10.48550/arXiv.1606.06565

ALPAYDIN, E. (2020). *Introduction to Machine Learning* (4th ed.). MIT Press.

BALTRUŠAITIS, T., AHUJA, C., & MORENCY, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2), 423-443. https://ieeexplore.ieee.org/abstract/document/8269806

BEKKUM, M. v., ZUIDERVEEN BORGESIUS, F. J. (2021), Digital welfare fraud detection and the Dutch SyRI judgment. *European Journal of Social Security,* 23 (4), 323-340. https://doi.org/10.1177/13882627211031257

BHUYAN, B. P., RAMDANE-CHERIF, A., TOMAR, R., & SINGH, T. P. (2024). Neuro-symbolic artificial intelligence: a survey. *Neural Computing and Applications*, 36, 12809–12844. https://link.springer.com/article/10.1007/s00521-024-09960-z

BRAUNE, J., (2019) The OCEAN Big Five Personality Traits.https://www.brandspeak.co.uk/blog/market-research-and-the-ocean-big-five-personality-traits/

BUOLAMWINI, J., GEBRU, T. (2018), Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research,* 81, 1-15. https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

CAÑABATE PÉREZ, J. (2024). La auditoría de algoritmos del sistema de justicia penal. En Vallespín Pérez, D. (Coord.), Asencio Gallego, J.M., Jiménez Cardona, N., *Algoritmización de la Justicia Penal,* (pp. 183 a 197). Juruá Editorial.

Deloitte (2024), *The Global Future of Cyber Survey, 4TH Edition*. Deloitte. https://www.deloitte.com/content/dam/assets-shared/docs/services/risk-advisory/2024/deloitte-global-future-of-cyber-survey-4th-edition-the-promise-of-cyber.pdf

DINAN, E., FAN, A., WILLIAMS, A., URBANEK, J., KIELA, D., & WESTON, J. (2019). "Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation. 1-14. https://doi.org/10.48550/arXiv.1911.03842

CRIADO PÉREZ, C. (2020). *Invisible Women. Exposing Data Bias in a World Designed for Men*. Chatto & Windus.

D'IGNAZIO C., KLEIN L. F. (2020). *Data Feminism*. The MIT Press.

EUBANKS, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor*. Picador, St Martin's Press.

Éticas, (2021), *Guía de auditoría algorítmica*. https://eticas.ai/wp-content/uploads/2024/04/Guide-to-Algorithmic-Auditing-SP.pdf

GOODFELLOW, I., BENGIO, Y., & COURVILLE, A. (2016). *Deep Learning*. MIT Press.

European Data Protection Supervisor-EDPS (2024), *TechSonar reports.* https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar_en

MOON, J., LEE, H. S., CHU, J., PARK, D., HONG, S., SEO, H., JEONG, D., KONG, S., HAM, M. (2024). A New Frontier of AI: On-Device AI Training and Personalization. En *2024 IEEE/ACM 46th International Conference on Software Engineering: Software Engineering in Practice (ICSESEIP)*, (pp. 323-333). IEEE Computer Society. https://dl.acm.org/doi/10.1145/3639477.3639716

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

LIANG, P. P., ZADEH, A., & MORENCY, L. P. (2024). Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56 (10), 1-42. https://dl.acm.org/doi/full/10.1145/3656580

RACHOVITSA, A., JOHANN, N., (2022), The Human Rights Implications of the Use of AI in the Digital Welfare State: Lessons Learned from the Dutch SyRI *Case, Human Rights Law Review,* 22 (2). https://doi.org/10.1093/hrlr/ngac010

ROSE, K., ELDRIDGE, S., & CHAPIN, L. (2015). *The Internet of Things: An Overview*. Internet Society.

RUSSELL, S., & NORVIG, P. (2020). *Artificial Intelligence: A Modern Approach (4th ed.)*. Pearson.

DHAR, S., GUO, J., LIU, J., TRIPATHI, S., KURUP, U., SHAH. M. (2021). A Survey of On-Device Machine Learning: An Algorithms and Learning Theory Perspective. *ACM Trans. Internet Things,* 2 (3), 1-49. https://arxiv.org/pdf/1911.00623

SMITH, G., & RUSTAGI, I. (2021). When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity. *Stanford Social Innovation Review*. https://doi.org/10.48558/A179-B138