Cómo citar: Coddou Mc Manus, Alberto, Mariana Germán Ortiz y Reinel Tabares Soto. 2025. "Evitando la trampa del formalismo: evaluación crítica y selección de métricas de equidad estadística en algoritmos públicos". *Revista de Estudios Sociales* 93: 85-106. https://doi.org/10.7440/res93.2025.05

Evitando la trampa del formalismo: evaluación crítica y selección de métricas de equidad estadística en algoritmos públicos*

Alberto Coddou Mc Manus, Mariana Germán Ortiz y Reinel Tabares Soto

Recibido: 18 de diciembre de 2024 | Aceptado: 6 de mayo de 2025 | Modificado: 3 de junio de 2025
https://doi.org/10.7440/res93.2025.05

Resumen | Este artículo analiza las distintas métricas de equidad estadística utilizadas para medir el desempeño de los modelos de inteligencia artificial (IA) y propone criterios para su aplicación según el contexto y las implicancias jurídicas involucradas. En particular, el trabajo examina cómo estas métricas pueden contribuir a garantizar el derecho a la igualdad y no discriminación en sistemas algorítmicos implementados por el Estado. La contribución central de este trabajo radica en la construcción de un marco analítico que permite seleccionar métricas de equidad estadística (*fairness*) según la finalidad del sistema automatizado, la naturaleza del proyecto y los derechos en juego. Por ejemplo, en el ámbito penal, donde está en riesgo la libertad individual, se enfatiza la necesidad de minimizar falsos positivos, mientras que en algoritmos destinados a proteger a víctimas de violencia de género se prioriza reducir falsos negativos. En contextos como la contratación pública, se evalúa la equidad grupal mediante métricas como el impacto dispar o la paridad demográfica, y en sectores como la fiscalización tributaria o el diagnóstico médico, se privilegian la precisión predictiva y la eficiencia. A partir de un enfoque interdisciplinario, el artículo propone una mirada sociotécnica que integra perspectivas técnicas y jurídicas. Se destaca la necesidad de evitar la "trampa del formalismo", consistente en que la equidad se reduce a métricas abstractas sin considerar el contexto social y político. Finalmente, se subraya que la adopción de métricas adecuadas no solo permite detectar y mitigar sesgos algorítmicos, sino también contribuir a la implementación de sistemas de IA más justos y transparentes, alineados con principios fundamentales de igualdad y no discriminación.

Palabras clave | algoritmos públicos; discriminación algorítmica; equidad estadística; inteligencia artificial; justicia algorítmica; sesgo algorítmico

Avoiding the Formalism Trap: A Critical Evaluation and Selection of Statistical Fairness Metrics in Public Algorithms

Abstract | This article examines the different statistical fairness metrics used to evaluate the performance of artificial intelligence (AI) models and proposes criteria for selecting them based on context and legal implications. It focuses in particular on how these metrics can help safeguard the right to equality and non-discrimination in algorithmic systems implemented by the state. Its core contribution is the development of an analytical framework for choosing fairness metrics according to the purpose of

* La investigación de la que deriva este artículo fue financiada por la Agencia Nacional de Investigación y Desarrollo (ANID), Chile: Subdirección de Investigación Aplicada / Beca IDeA I+D 2023 (folio ID23I10357), proyecto Fondecyt regular 1230895 y proyecto Fondecyt de iniciación 11220370. Además, contó con el apoyo del GobLab de la Universidad Adolfo Ibáñez, Chile. Los autores y la autora de este artículo contribuyeron al desarrollo de la investigación y a la preparación, escritura y edición del texto.

the automated system, the nature of the project, and the rights at stake. For example, in the criminal justice system—where individual liberty is at risk—the emphasis is on minimizing false positives. In contrast, for algorithms designed to protect victims of gender-based violence, the priority is to reduce false negatives. In areas like public procurement, group fairness is assessed using metrics such as disparate impact or demographic parity. In sectors like tax enforcement or medical diagnostics, the focus is on predictive accuracy and efficiency. Taking an interdisciplinary approach, the article puts forward a sociotechnical perspective that brings together technical and legal insights. It highlights the need to avoid the "formalism trap," where fairness is reduced to abstract metrics without accounting for the broader social and political context. Finally, it argues that selecting appropriate metrics not only helps identify and mitigate algorithmic bias but also contributes to building AI systems that are fairer and more transparent, and that align with fundamental principles of equality and non-discrimination.

Keywords | algorithmic bias; algorithmic discrimination; algorithmic justice; artificial intelligence; public algorithms; statistical fairness

Evitando a armadilha do formalismo: avaliação crítica e seleção de métricas de equidade estatística em algoritmos públicos

Resumo | Neste artigo, analisam-se as diferentes métricas estatísticas de equidade utilizadas para medir o desempenho de modelos de inteligência artificial (IA) e se propõem critérios para sua aplicação de acordo com o contexto e com as implicações legais envolvidas. Em particular, examina-se como essas métricas podem contribuir para garantir o direito à igualdade e à não discriminação nos sistemas algorítmicos implementados pelo Estado. A contribuição central deste trabalho reside na construção de um quadro analítico que permita a seleção de métricas de equidade estatística (*fairness*) de acordo com a finalidade do sistema automatizado, com a natureza do projeto e com os direitos envolvidos. Por exemplo, na esfera criminal, em que a liberdade individual está em risco, enfatiza-se a necessidade de minimizar os falsos positivos, enquanto os algoritmos voltados para a proteção das vítimas de violência de gênero priorizam a redução dos falsos negativos. Em contextos como a contratação pública, avalia-se a equidade do grupo usando métricas como impacto díspar ou paridade demográfica, e em setores como auditoria fiscal ou diagnóstico médico, priorizam-se a precisão preditiva e a eficiência. A partir de uma abordagem interdisciplinar, o artigo propõe uma perspectiva sociotécnica que integra perspectivas técnicas e jurídicas. Destaca-se a necessidade de evitar a "armadilha do formalismo", que consiste em reduzir a equidade a métricas abstratas sem considerar o contexto social e político. Por fim, ressalta-se que a adoção de métricas adequadas não apenas permite detectar e mitigar vieses algorítmicos, mas também contribui para a implementação de sistemas de IA mais justos e transparentes, alinhados aos princípios fundamentais de igualdade e de não discriminação.

Palavras-chave | algoritmos públicos; discriminação algorítmica; equidade estatística; inteligência artificial; justiça algorítmica; viés algorítmico

Introducción

En la era digital, el rápido avance de la inteligencia artificial (IA) y el aprendizaje automático han revolucionado múltiples sectores, desde la salud hasta el sistema financiero. Sin embargo, este progreso ha traído consigo una serie de desafíos éticos y sociales, y uno de los más críticos es garantizar que tanto los procedimientos como los resultados de los sistemas automatizados de decisión cumplan con estándares de justicia. En este escenario, una de las cuestiones más relevantes es asegurar que los sistemas automatizados que utilizan herramientas de IA estén libres de sesgos, cumplan con normas básicas de equidad o imparcialidad, y que no infrinjan estándares de igualdad y no discriminación.

Sin embargo, la literatura sobre los impactos éticos y jurídicos de la IA usa expresiones con diversos significados, dependiendo de una multiplicidad de factores. Así, los términos *sesgo*, *equidad*, *imparcialidad* y *discriminación* suelen tener varios sentidos en diversas publicaciones científicas, cuestión que dificulta la posibilidad de lograr un consenso técnico que permita evaluar el avance, la utilidad o el impacto de las investigaciones.

Si bien el término *equidad* se puede interpretar de distintas maneras, se suele asociar con el estándar de un trato justo y equitativo o imparcial. Por ejemplo, de acuerdo con el trabajo de Rawls *Justicia como equidad* (2012), la equidad es una característica de la estructura de las instituciones políticas y sociales, que deriva no de una concepción moral, filosófica o religiosa, sino de cierto tipo de procedimiento que nos permite adoptar arreglos institucionales comunes que puedan ser calificados como justos. En otros escenarios, en cambio, la equidad es asociada a la idea de un trato no arbitrario, como sucede en diversas jurisdicciones; es el caso del ámbito de las disputas sobre inversiones extranjeras, en el que la equidad es entendida como un trato "justo y equitativo", cuestión que depende de verificar la existencia de razones objetivas e imparciales que permitan justificar un determinado trato y la ausencia de animadversión o irracionalidad en la decisión. En ciertas jurisdicciones, este estándar se asocia con el cumplimiento de alguna norma sustantiva, como aquellas que forman parte del derecho de la antidiscriminación (Vandevelde 2010). En este trabajo, cuando usemos el término *equidad estadística* (*statistical fairness*), nos referiremos a la idea de medir la equidad a partir de criterios objetivos, de métricas que permitan decir si un determinado sistema automatizado cumple con cierto estándar de justicia previamente establecido. Por su parte, cuando hablemos de *sesgos*, haremos referencia al modo en que la literatura sobre psicología cognitiva se ha ido adaptando al estudio de procesos de toma de decisión para dar cuenta de errores o problemas que afectan esos procesos. Por último, emplearemos el término *discriminación* de acuerdo con concepciones jurídicas relativamente asentadas en el derecho internacional y en diversas jurisdicciones que han desarrollado ciertas categorías doctrinarias, como la discriminación directa, la discriminación indirecta y la obligación de realizar ajustes razonables (Moreau 2010).

Actualmente, la literatura en ciencia de datos ha identificado diversas formas de medir el sesgo de los modelos de IA. Entre ellas, se destacan las métricas estadísticamente centradas, como la precisión, la tasa de falsos positivos o la paridad demográfica (Verma y Rubin 2018). Si bien algunas de estas métricas son fáciles de aplicar, como el test de precisión, existe una falta de reflexión crítica acerca del modo en que pueden evitar la evaluación de eventuales infracciones al derecho a la igualdad y a la no discriminación. Como sostienen Wachter, Mittelstadt y Russell (2020), el intenso debate regulatorio que caracteriza el escenario actual exige una reflexión más acabada acerca de cómo las métricas desarrolladas en las ciencias de la computación pueden ser utilizadas en juicios de responsabilidad por infracción a estándares derivados del derecho a la igualdad y a la no discriminación.

El propósito de este artículo es describir y analizar las diferentes métricas de equidad estadística que se usan para medir el desempeño de los modelos, mapearlas según el tipo de proyecto o modelo de IA y evaluar críticamente su utilización de acuerdo con estándares derivados del derecho a la igualdad y a la no discriminación. En concreto, analizaremos la aplicabilidad de distintas métricas de *fairness* en distintos contextos, ofreciendo ejemplos relacionados con proyectos de IA desarrollados por el Estado chileno, e identificando qué métrica de equidad estadística es más adecuada según el tipo de modelo y el contexto en el que se implementa. Los ejemplos y casos concretos son extraídos del repositorio de Algoritmos Públicos de la Universidad Adolfo Ibáñez (Hermosilla y Germán 2024) o de otros casos de uso de sistemas de decisiones automatizadas utilizadas por el Estado de Chile.

Este análisis permitirá observar cómo las métricas de *fairness* se ajustan a proyectos públicos con distintos objetivos, desde políticas sociales hasta aplicaciones en seguridad o salud, o en el ámbito del empleo público. A partir de estos casos, se elaborará un bosquejo inicial de tipologías de métricas de *fairness*, que permitirá categorizarlas según la naturaleza del proyecto y las implicancias éticas y legales de cada uno. En tal sentido, la contribución de este trabajo es ofrecer algunas justificaciones para la utilización de métricas de equidad estadística que permitan aproximarse al cumplimiento de los estándares derivados del derecho antidiscriminación. Creemos que la interdisciplinariedad del equipo de investigación, que integra a las ciencias jurídicas con las ciencias de la computación, es un reflejo de nuestra intención de contribuir a un debate complejo que, a veces, está sujeto a presiones por ofrecer respuestas rápidas y certeras, pero escasamente reflexionadas. En este contexto, y tal como señala el título de este texto, evitar la "trampa del formalismo" supone realizar un esfuerzo por profundizar en conceptos como *sesgo*, *equidad*, *justicia* y *no discriminación*, cuyos significados pueden ser procedimentales o contextuales, pero que, por sobre todo, están sujetos a discusión, de modo que los conflictos no se resuelven mediante formalismos matemáticos (Selbst *et al*. 2019). Evitar la trampa del formalismo implica ir más allá, hacia un ejercicio que puede resultar especulativo para algunos, pero que ofrece razones para acercar dos mundos que, hasta el momento, parecen estar sumidos en un diálogo poco productivo.

La distinción entre sesgo y discriminación

En la ley de inteligencia artificial de la Unión Europea (UE) (Regulation [EU] 2024/1689) se establece una serie de obligaciones para los proveedores de sistemas de IA que sean calificados como de alto riesgo. Entre estas obligaciones, destacan aquellas relativas a la gobernanza de datos y, en especial, todo lo relacionado con el principio de calidad de los datos. El artículo 10 de esta normativa, que entró en vigencia en agosto de 2024, establece una serie de criterios de calidad que deben cumplirse respecto a los conjuntos de datos de "entrenamiento, validación y prueba" de los sistemas de IA. El artículo señala que, entre las prácticas de gobernanza y gestión de datos, se consideran las cuestiones relativas al diseño, la recolección y la finalidad de los datos; las "operaciones de tratamiento oportunas para la preparación de los datos, como la anotación, el etiquetado, la depuración, la actualización, el enriquecimiento y la agregación"; la información que los datos dicen medir; una "evaluación de la disponibilidad, la cantidad y la adecuación de los conjuntos de datos necesarios", y, quizás lo más importante, la evaluación de los "posibles sesgos que puedan afectar a la salud y la seguridad de las personas, afectar negativamente a los derechos fundamentales o dar lugar a algún tipo de discriminación prohibida por el Derecho de la Unión, especialmente cuando las salidas de datos influyan en las informaciones de entrada de futuras operaciones". Ante la existencia de estos posibles sesgos, los proveedores de sistemas de IA deben adoptar las medidas necesarias para detectarlos, prevenirlos y mitigarlos (Bekkum 2024).

De la lectura de esta disposición, destaca el hecho de que se utilizan dos términos que, a veces, en el marco de los debates sobre los efectos negativos que pueden tener las tecnologías digitales (y, en especial, la IA), se entienden como sinónimos: sesgo y discriminación. Una interpretación literal del artículo 10 de la Ley de IA de la UE nos permite sostener que un sistema que asegure estar libre de sesgos puede, en términos normativos, infringir las normas del derecho de la antidiscriminación que están vigentes en el derecho de la UE. En otras palabras, se trataría de cuestiones distintas: un proveedor de sistemas de IA puede perfectamente desplegar sus mejores esfuerzos por detectar, prevenir y mitigar posibles sesgos, pero ello no asegura que el sistema esté libre de discriminación, al menos en los términos utilizados por una definición estándar de este derecho fundamental.

Esta distinción entre sesgo y discriminación es especialmente relevante en contextos regulatorios, ya que implica que el cumplimiento técnico con criterios de calidad de datos, por ejemplo, mediante la detección y corrección de sesgos estadísticos o algorítmicos, no garantiza por sí solo el cumplimiento jurídico con respecto al derecho antidiscriminación. Por ello, más que oponer lo técnico y lo jurídico, es necesario entender que el análisis técnico de sesgos debe guiarse por el contexto del problema que se busca resolver, considerando las consecuencias sociales y normativas asociadas. Esto implica seleccionar y aplicar métricas de equidad adecuadas a cada caso, integrando dimensiones como la distribución del error, el impacto sobre grupos protegidos y la finalidad del sistema. Es precisamente en esta articulación entre el enfoque técnico y las exigencias del derecho en la que el análisis legal resulta clave, no solo para interpretar adecuadamente los riesgos, sino también para orientar las decisiones metodológicas sobre qué medir, cómo medirlo y con qué criterios evaluar si un sistema ofrece razones para ser calificado como acorde con el derecho.

Este ejemplo da cuenta de la necesidad de abordar el uso y la comprensión de los términos *sesgo* y *discriminación* de manera conjunta, sobre todo si los países de la región latinoamericana están comenzando a importar la legislación europea y a adaptarla a sus propias realidades. Tal como sucedió con el impacto regional que tuvo el Reglamento Europeo de Protección de Datos, es esperable que la ley de IA de la UE tenga diversos efectos sobre los ordenamientos jurídicos latinoamericanos (Contreras 2024). A comienzos de 2024, e inspirado en la regulación europea, el Gobierno de Chile, a través del Ministerio de Ciencia, Tecnología, Conocimiento e Innovación, presentó un proyecto de ley que regula los sistemas de IA. En este proyecto, se incluye un artículo especial que contiene una serie de principios aplicables a los sistemas de IA, dentro de los que se destaca el de "diversidad, no discriminación y equidad", que exige que estos sistemas se desarrollen y utilicen "durante todo su ciclo de vida, promoviendo la igualdad de acceso, la igualdad de género y la diversidad cultural, evitando al mismo tiempo los efectos discriminatorios y sesgos de selección o de información que pudieran generar un efecto discriminatorio" (artículo 4, e). Nuevamente, las fuentes normativas parecen no esclarecer la aparente confusión que existe entre dos términos que exigen un tratamiento y una reflexión más detallados.

El término sesgo proviene principalmente de las matemáticas, específicamente de la estadística. De acuerdo con la definición del *Diccionario de la lengua española* de la RAE (2025), un sesgo es un "error sistemático en el que se puede incurrir cuando al hacer muestreos o ensayos se seleccionan o favorecen unas respuestas frente a otras". Este término, valorativamente neutro, se usa para describir aquellas situaciones que pueden ocurrir al momento de diseñar, implementar o testear un sistema de IA que funciona sobre la base de un conjunto de datos. En las ciencias de la computación, a veces el término se utiliza de esta manera, como un mero error en el que es posible incurrir y que no compromete la responsabilidad del proveedor de sistemas de IA. Sin embargo, cuando la preocupación es acerca de la calidad o utilidad de los sistemas de IA para la adopción de decisiones que tradicionalmente eran adoptadas por seres humanos y que pueden tener efectos jurídicos, lo que se busca es precisamente evitar errores que puedan no solo comprometer la eficiencia o utilidad de estos procesos, sino el cumplimiento de ciertos estándares de justicia. De este modo, si lo que se quiere es garantizar que los sistemas computacionales de IA estén basados en conjuntos de datos validados y testeados de un modo que permita asegurar un mínimo de calidad, es inevitable pensar que existe cierto tipo de sesgos que dejan ver la responsabilidad de los proveedores por infringir sus obligaciones en materia de igualdad y no discriminación.

Asegurar la calidad de los datos, desde una perspectiva técnica, implica velar por la integridad y robustez de los datos que alimentan estos sistemas (Palma 2024). Esto incluye aspectos como la completitud, la ausencia de valores nulos o erróneos, la accesibilidad de las variables necesarias, el equilibrio entre las clases o categorías representadas, y la

detección y el tratamiento de valores atípicos que puedan distorsionar los resultados. Estos criterios son fundamentales para garantizar que el sistema aprenda de manera adecuada y no genere errores sistemáticos que afecten su funcionamiento o legitimidad. Sin embargo, el concepto de calidad, al estar incorporado en una norma jurídica sobre las obligaciones relativas a la gobernanza de datos, tal como se deriva del artículo 10 de la ley de IA de la UE, supone no solamente asegurarse de que sean precisos y pertinentes, sino evaluar el modo en que estos y los criterios para su análisis o procesamiento dan cuenta de sistemas o procesos que previenen violaciones al derecho a la igualdad y no discriminación o a otros derechos fundamentales. En este sentido, un proceso o sistema automatizado de decisión técnicamente adecuado es uno que también considera eventuales infracciones a las normas sobre derechos fundamentales (Palma 2024).

Las personas dedicadas a las ciencias de la computación, y específicamente quienes investigan el diseño e implementación de sistemas de IA, suelen hablar de métricas de equidad (*fairness*) para dar cuenta de modelos que estén libres de sesgos y, en tal sentido, garantizar el cumplimiento de ciertos estándares de justicia. Sin embargo, estas métricas operan bajo un concepto de radical "abstracción": los sistemas pueden describirse como "cajas negras", definidas por sus entradas, salidas y la relación entre ellas (Selbst *et al*. 2019). De este modo, las propiedades deseadas de un sistema pueden describirse solo en términos de entradas y salidas, sin considerar la procedencia de estas ni los detalles internos del sistema.

Los sistemas de IA se diseñan y construyen para alcanzar objetivos y métricas de rendimiento específicos (precisión, *recall*, paridad demográfica, etc.). En este escenario, ha surgido el campo del aprendizaje automático consciente de la equidad (*fair-ML* [*machine learning*]), que busca diseñar algoritmos y modelos de ML más justos, utilizando la equidad como una propiedad del sistema (Caton y Haas 2024). La literatura en este ámbito es profusa, y sobre todo aplica diversos criterios de equidad para la evaluación de alguna noción de justicia algorítmica (Verma y Rubin 2018). Sin embargo, tal como sostienen Selbst *et al*. (2019), casi todos estos artículos definen el sistema de interés de forma limitada, circunscribiéndolo al modelo de aprendizaje automático, junto a sus entradas y salidas, y abstrayendo cualquier contexto que rodee a este sistema. Así, los componentes centrales de la equidad algorítmica son el desarrollo de definiciones matemáticas para una toma de decisiones que pueda ser calificada como "justa", la optimización de algoritmos en función de estas definiciones y la auditoría de algoritmos para detectar posibles violaciones de estas definiciones (Green 2022).

Sin embargo, el problema es doble. En primer lugar, incluso cuando estamos frente a sistemas que declaran estar libres de sesgos, al menos con respecto a ciertas métricas, los tomadores de decisiones pueden incurrir en una suerte de *aversión algorítmica*, es decir, un tipo de sesgo que se caracteriza por su contraste con el denominado sesgo de automatización (Chacón, Kausel y Reyes 2022). En estos casos, el derecho o garantía a la supervisión humana sobre la IA podría incluso generar una suerte de desconfianza en los sistemas de *fair-ML*, lo que disminuiría el potencial de la IA para promover mayor igualdad de oportunidades. Por otra parte, los problemas surgen cuando se pretende traducir estas definiciones o métricas al lenguaje del derecho (Wachter, Mittelstadt y Russell 2020).

Un enfoque común que se utiliza para dar cuenta de que un determinado algoritmo cumple con el derecho de la antidiscriminación es analizar las diferencias de trato entre grupos protegidos y no protegidos. Sin embargo, existen diversas maneras de medir estas diferencias. Una de las más utilizadas es la *paridad estadística/demográfica*, que examina el porcentaje total de asignaciones o clasificaciones positivas o negativas entre los grupos. No obstante, esta métrica es limitada, ya que no distingue entre casos de discriminación que podrían estar justificados por algún interés legítimo. Por ejemplo, si se exigiera paridad estadística para que las tasas de predicción positiva, es decir, las predicciones de alta probabilidad de reincidencia, fueran iguales para hombres y mujeres en sistemas

de evaluación de riesgo penal, se podría incurrir en una injusticia: las mujeres, que estadísticamente tienen menores tasas reales de reincidencia, podrían ser clasificadas erróneamente como de alto riesgo solo para igualar las tasas con las de los hombres, lo que resultaría en que pasen más tiempo en prisión sin justificación basada en evidencia (Green 2022).

En este contexto, conceptos como los de sesgo y equidad están arraigados en disciplinas como la estadística y la informática, y recientemente se han expandido hacia el campo de la ética aplicada a la IA, y se ha generado una activa discusión en torno a los desafíos que enfrenta el actual desarrollo tecnológico. Sin embargo, los significados que estos términos adquieren en los campos técnico y ético no siempre coinciden con la forma en que se entienden o abordan los problemas jurídicos relacionados con la igualdad y la no discriminación. El sesgo abarca mucho más que la discriminación, pues incluye un amplio rango de errores sistemáticos, que pueden ser de tipo estadístico, cognitivo, social, estructural o institucional, y no solo aquellos que generan resultados injustos. En el caso de la *equidad algorítmica*, el *sesgo algorítmico* describe errores que sistemáticamente favorecen a grupos privilegiados y desfavorecen a otros. Aunque existe cierto solapamiento con la noción legal de *discriminación*, el sesgo algorítmico es más amplio porque abarca cualquier desventaja considerada ética o moralmente problemática. Desde una perspectiva legal, en cambio, la *discriminación algorítmica* se refiere específicamente a un trato injusto o perjudicial hacia individuos de grupos protegidos, ya sea explícitamente definidos por la ley (por ejemplo, por características específicas) o implícitamente reconocidos a través de un proceso de reconstrucción dogmática del derecho a la antidiscriminación (Martínez Placencia 2023).

Diferentes tipos de sesgos en un algoritmo público

Siguiendo la revisión de literatura desarrollada por Van Giffen, Herhausen y Fahse (2022), quienes además propusieron una taxonomía de los diferentes tipos de sesgos y sus posibles medidas de mitigación, esta sección pretende ilustrar, a través del análisis de un algoritmo público, los problemas que pueden suscitarse en las diferentes fases del ciclo de vida de un algoritmo. En concreto, describimos un algoritmo público desarrollado por el Ministerio Público chileno que tiene por objeto la automatización de la evaluación del riesgo de las mujeres víctimas de violencia en relaciones de pareja (Coddou *et al*., en prensa)[1]. Implementado inicialmente en cinco fiscalías regionales, el algoritmo de sugerencia de riesgo (en adelante, ASR) sirve como una herramienta para apoyar la toma de decisiones de las fiscalías, alertando al personal cuando es necesario modificar las medidas de protección para la víctima y, eventualmente, solicitar medidas cautelares para los imputados, que deben ser finalmente decretadas por un juez. Este sistema se ha utilizado en más de 8.000 casos, enfocándose específicamente en delitos de violencia intrafamiliar (VIF).

El desarrollo del ASR busca mejorar la evaluación de riesgo frente a la violencia de género, respondiendo a los problemas que actualmente presenta la Pauta Unificada de Evaluación Inicial de Riesgo (en adelante, PUIR), un cuestionario de preguntas cerradas que no se actualiza en el tiempo, que mide algunas variables de apreciación subjetiva y que, en general, es implementado por personal no especializado. El propósito declarado del ASR es proporcionar información de calidad y en tiempo real para optimizar la toma de decisiones relacionadas con la protección de víctimas. Las diferentes unidades de atención a las víctimas y testigos de cada fiscalía (Uravit) utilizan el ASR para diseñar estrategias de protección, evaluando el riesgo de revictimización mediante un análisis de datos de casos de violencia intrafamiliar desde 2015. El algoritmo se actualiza mensualmente para

1 Incluido en el repositorio de Algoritmos Públicos elaborado por el GobLab de la Universidad Adolfo Ibáñez. Véase https://www.algoritmospublicos.cl/proyecto-Algoritmo-Sugerencia-Riesgo-Min-Publico

mejorar sus capacidades predictivas y permite emitir alertas a las Uravit y a fiscales en caso de riesgo elevado. En concreto, el ASR emplea veintiséis variables principales y otras adicionales para evaluar el riesgo. Entre las variables se incluyen antecedentes de violencia de parte del agresor y datos específicos de la víctima, como historial de denuncias y el cumplimiento de medidas de protección. El algoritmo también emplea un modelo de procesamiento de lenguaje natural que identifica patrones en los textos de denuncias y declaraciones. La tecnología utilizada permite un análisis rápido de grandes volúmenes de datos y actualizaciones periódicas para ajustar la situación de riesgo de cada víctima.

Para nuestro análisis usaremos la taxonomía de los diferentes tipos de sesgos (Giffen, Herhausen y Fahse 2022) que pueden afectar a un sistema de IA de aprendizaje automatizado a partir del ejemplo dado en el párrafo anterior.

1. Sesgo social: Este primer tipo de sesgo supone que los datos que se recolectan de la población relevante pueden estar afectados por una serie de estereotipos o prejuicios sociales. Así, por ejemplo, si existe un prejuicio social de que la violencia económica no es una forma de violencia en relaciones de pareja, entonces los datos históricos que constituyen la base del ASR no consideran este tipo de violencia como un factor en la evaluación del riesgo.

2. Sesgo de representación: Si el conjunto de datos de entrenamiento se constituye con base en los datos de los sistemas de asistencia social (en el caso chileno, por ejemplo, el Registro Social de Hogares), el ASR podría ignorar u otorgar puntuaciones más bajas a las mujeres que no están representadas en tales registros, como las víctimas de violencia de pareja que pertenecen a sectores sociales que no dependen de la asistencia social. Esto suele suceder en algunos algoritmos de violencia de género a nivel comparado (por ejemplo, el caso de VioGen en España).

3. Sesgo de etiquetado: Este tipo de sesgo considera que gran parte de los conjuntos de datos de entrenamiento o de retroalimentación de los algoritmos están etiquetados o clasificados por seres humanos. Así, por ejemplo, los datos históricos de los casos de VIF contienen las apreciaciones o clasificaciones que realizaron los asistentes sociales en causas registradas ante la Fiscalía Nacional y que pudieron haber modificado los resultados de la aplicación de la pauta PUIR.

4. Sesgo de medición: Las variables seleccionadas pueden no representar el riesgo real si son *proxies* inadecuados para dar cuenta de los riesgos. Así, por ejemplo, si bien la literatura ha indicado que la percepción del miedo de la víctima es uno de los principales predictores de violencia, los instrumentos de medición pueden cometer el error de basarse únicamente en el autorreporte. De este modo, solo aquellas víctimas que serían más propensas a expresar o manifestar ante las autoridades policiales sus percepciones de miedo serían adecuadamente evaluadas por el ASR. Aquí, claramente, el problema reside en el instrumento.

5. Sesgo algorítmico: En ciencias de la computación, este parece ser el sesgo más genérico, y se define como aquel que ocurre cuando consideraciones técnicas inapropiadas durante la fase de diseño derivan en desviaciones sistemáticas de los resultados. De este modo, si el modelo no se ajusta para equilibrar los falsos positivos y negativos, podría errar en su categorización de riesgo. Así, por ejemplo, un ajuste que priorice evitar falsos negativos, es decir, que prefiera pecar por exceso antes que por defecto, por el riesgo que implica no proteger a una víctima que realmente necesita una medida de protección, podría sobreclasificar como de riesgo alto a algunas mujeres de minorías visibles, lo que afectaría el acceso igualitario a recursos.

6. Sesgo de evaluación: El algoritmo debe evaluarse en conjuntos de datos representativos. Así, si el modelo es testeado o sometido a pruebas de validación con datos de zonas urbanas, podría tener un rendimiento deficiente en áreas rurales al clasificar los niveles de riesgo en esas zonas. Lo mismo sucede si acaso los parámetros de comparación del rendimiento o justicia algorítmica, es decir, los *benchmarks*, están afectados a su vez por sesgos.

7. Sesgo de despliegue: Si un algoritmo se usa en un contexto diferente para el que fue diseñado, puede fallar. De este modo, si los resultados de la aplicación del ASR, que se utiliza para la evaluación de las víctimas de violencia de pareja, se emplean para determinar decisiones judiciales como la custodia de hijos, sin que la aplicación esté calibrada para ello, esto llevaría a consecuencias injustas.

8. Sesgo de retroalimentación: Las clasificaciones de riesgo derivadas de la aplicación del ASR pueden reforzar el sesgo en el futuro. De esta manera, si el algoritmo clasifica consistentemente como de alto riesgo a ciertas comunidades, podría aumentar la vigilancia en esas áreas, lo que generaría más reportes de incidentes y reforzaría el sesgo en futuras predicciones, y podría terminar afectando tanto a las víctimas como a los imputados por estos hechos de violencia.

El desarrollo de un modelo de aprendizaje automático puede introducir sesgos en varias etapas del ciclo de vida del algoritmo, desde la recopilación de datos hasta su implementación. En la recopilación de datos, es posible que los sesgos surjan por muestras no representativas o por procesos de selección que favorecen a ciertos grupos. En la definición del problema y la selección de características, las decisiones sobre los objetivos del modelo y las variables incluidas reflejarían desigualdades preexistentes. Durante el entrenamiento, los sesgos pueden provenir del algoritmo o de la configuración de sus parámetros. En la evaluación, el uso de métricas inadecuadas ocultaría impactos desiguales en grupos específicos. Finalmente, en la implementación, el uso del modelo reforzaría desigualdades, ya sea a través de decisiones basadas únicamente en los resultados del algoritmo o mediante ciclos de retroalimentación que perpetúan la discriminación. Abordar estos desafíos requiere una perspectiva integral y la colaboración de múltiples actores para garantizar la equidad en cada etapa.

Definición del problema: equidad y sesgo en IA

El concepto de equidad estadística en IA es amplio, pues abarca diferentes medidas para evaluar un algoritmo de acuerdo con parámetros o métricas objetivas que permiten la comparación de rendimientos algorítmicos. Es justamente este tipo de evaluaciones el que posibilitaría tomar decisiones sobre la base de algoritmos más o menos justos, al menos de acuerdo con estas métricas de equidad estadística. Tal vez el tema que ha generado más debate en la literatura sobre aprendizaje automático justo es cómo definir la equidad (Selbst *et al*. 2019). Dado que los algoritmos "hablan en el lenguaje de las matemáticas", la tarea principal dentro de quienes se dedican al *fair-ML* ha sido traducir las nociones inherentemente vagas de equidad o justicia social a definiciones matemáticas claras, para así integrar estos ideales en el diseño de los sistemas de aprendizaje automático que aspiran a ser operativos.

Este fenómeno, conocido como la *trampa del formalismo*, supone analizar todos los problemas que surgen al intentar definir la justicia algorítmica únicamente a través de métricas estadísticas. En este sentido, es crucial considerar el contexto social, histórico y político del problema, así como las implicaciones del algoritmo para diferentes grupos e individuos. De este modo, la equidad estadística es solo una herramienta para evaluar la justicia algorítmica, y no debe ser utilizada como un sustituto del análisis crítico y la

reflexión ética. No obstante, es importante reconocer que, incluso dentro de los enfoques que buscan formalizar la equidad mediante métricas estadísticas, existen profundas discrepancias respecto a cómo abordar de forma adecuada lo que a menudo se denomina "el problema formal". A pesar de los avances en la definición y aplicación de distintas métricas de detección de sesgos, la comunidad académica aún no ha alcanzado un consenso sobre cuáles de estas métricas deben priorizarse en cada caso ni sobre cómo resolver las tensiones inherentes entre ellas. Esta falta de acuerdo no solo evidencia la complejidad del problema, sino que refuerza la necesidad de enfoques interdisciplinares que combinen lo técnico, lo ético y lo jurídico para guiar la toma de decisiones en torno a la justicia algorítmica.

Las ciencias de la computación contribuyen con su conocimiento técnico para mitigar sesgos en los datos y los modelos, aunque deben evitar el formalismo de centrarse solo en métricas matemáticas sin considerar el contexto social. Las ciencias jurídicas ofrecen el marco normativo para garantizar la no discriminación y los derechos fundamentales, interpretando la legislación en contextos nuevos, como el de la inteligencia artificial (Ungern-Sternberg 2022). Las ciencias sociales, por su parte, analizan los sesgos históricos, sociales y culturales que se reflejan en los algoritmos, así como su impacto en la sociedad. Juntas, estas disciplinas permiten diseñar algoritmos más justos, desarrollar marcos de gobernanza efectivos y fomentar un debate público informado, integrando valores técnicos, legales y sociales para garantizar estándares de justicia adecuados a la era de la IA.

Métricas de *fairness*: clasificación y desafíos

La evaluación del *fairness* en los modelos de IA se realiza a través de métricas estadísticas que permiten medir la equidad de los resultados en términos de distribución entre diferentes grupos poblacionales. A continuación, describimos métricas comúnmente utilizadas para medir sesgo en la aplicación de *fair-ML*, ofreciendo algunos ejemplos que nos permiten ilustrar qué es lo que cada una de ellas aspira a certificar.

1. Paridad demográfica: Esta métrica se centra en asegurar que la proporción de resultados positivos sea igual para todos los grupos demográficos, independientemente de sus atributos. El objetivo es que un algoritmo no favorezca o perjudique a un grupo en particular. En otras palabras, busca garantizar que, por ejemplo, hombres y mujeres o diferentes grupos étnicos tengan la misma probabilidad de recibir un resultado favorable (como la aprobación de un préstamo), independientemente de otras características.

   Ejemplo: En un modelo de concesión de préstamos, si el 60% de los hombres recibe aprobación y solo el 40% de las mujeres, entonces se estaría violando la paridad demográfica. Para cumplir esta métrica, el porcentaje de aprobaciones debería ser el mismo entre hombres y mujeres, sin importar si sus características financieras son comparables o no. Es importante destacar que esta definición no implica que todos los grupos tengan igual desempeño financiero promedio, sino que todos reciban resultados positivos en igual proporción, lo que a veces puede generar tensiones con otras métricas de equidad que sí consideran mérito o necesidad.

2. Paridad de resultados: También conocida como *equidad de oportunidad*, esta métrica se enfoca en asegurar que los resultados del modelo sean equitativos en términos de los errores que comete, en lugar de simplemente los aciertos. Aquí, el objetivo es que la tasa de falsos positivos y falsos negativos sea similar para diferentes grupos demográficos. También pueden usarse métricas como la tasa de falsa omisión y la tasa de falso descubrimiento. La paridad de resultados es particularmente importante

cuando los errores de un modelo pueden tener consecuencias desproporcionadas para ciertos grupos y afectan de manera distinta dependiendo de si son métricas basadas en falsos positivos o en falsos negativos.

> *Ejemplo*: En un modelo de reconocimiento facial utilizado por fuerzas de seguridad, la paridad de resultados significaría que la tasa de falsos positivos (identificaciones incorrectas de personas) es la misma para personas de diferentes etnias. Si el sistema identifica erróneamente a personas de una etnia específica con mayor frecuencia, estaría violando esta métrica, incluso si su precisión general es alta.

3. Impacto dispar: Esta métrica se utiliza para evaluar si un modelo tiene un impacto desproporcionado en diferentes grupos, incluso cuando los resultados del modelo no son directamente discriminatorios. El concepto de impacto dispar se originó en el derecho antidiscriminación y busca identificar decisiones que, aunque aparentemente neutrales, afectan negativamente a grupos protegidos. La métrica evalúa la tasa entre la proporción de resultados positivos entre grupos desfavorecidos y favorecidos, y se considera un problema cuando esa ratio *cae* por debajo de un umbral predefinido. En algunas jurisdicciones, este ha sido el criterio para determinar presunciones de discriminación, tal como sucede en el caso de la denominada regla de los cuatro quintos del derecho laboral estadounidense. En términos de *fairness*, el impacto dispar mide la equidad de un proceso evaluando la proporción de resultados favorables entre los grupos. Si la tasa de éxito o selección de un grupo es significativamente menor que la del grupo de referencia (generalmente, por debajo del 80%, como en el caso del derecho laboral estadounidense), puede considerarse que el grupo está experimentando un impacto adverso. Esta métrica no necesariamente considera las razones por las cuales ocurre la disparidad, sino que se enfoca en los resultados observados y en la necesidad de alcanzar una proporción más equitativa entre los grupos evaluados.

   > *Ejemplo*: Un sistema de puntuación crediticia que no utiliza la raza como variable directa podría, sin embargo, tener un impacto dispar si las variables que emplea, como el código postal o la ocupación, están correlacionadas con la raza. Aunque el modelo no discrimine explícitamente, si el resultado es que las personas de ciertos grupos raciales reciben menos préstamos que otros, entonces podría estar produciendo un impacto dispar, violando así este principio de equidad.

4. Justicia individual: A diferencia de las métricas anteriores, que se enfocan en resultados a nivel de grupo, la justicia individual o la precisión predictiva evalúa si los individuos que son similares entre sí en términos de características relevantes reciben el mismo tratamiento o resultado. Esta métrica es fundamental en contextos donde el resultado debe ser el mismo para casos análogos, independientemente de atributos que no deberían influir, como el género, la etnia o el lugar de origen. En esencia, busca garantizar que no haya diferencias de trato entre individuos que tendrían que ser considerados de manera equivalente.

   > *Ejemplo*: En el contexto del diagnóstico médico automatizado, la justicia individual requeriría que dos pacientes con síntomas y perfiles clínicos muy similares reciban el mismo diagnóstico y tratamiento recomendado, sin importar su género o etnia. Si el modelo produce un diagnóstico diferente para dos pacientes con características médicas idénticas debido a una variable irrelevante (como el género), violaría esta métrica.

Cada una de estas métricas tiene sus propias ventajas y limitaciones, lo que las hace más o menos adecuadas dependiendo del tipo de proyecto. Mientras la paridad demográfica es útil en proyectos de políticas públicas, puede no ser adecuada para modelos médicos, en los que la precisión individual es fundamental. La elección de una métrica sobre otra dependerá del contexto del proyecto, los valores de la organización y el impacto esperado sobre las personas afectadas por las decisiones algorítmicas.

Los principales desafíos y limitaciones que deben enfrentar las métricas de equidad estadística

Las métricas de *fairness* han sido aplicadas en una variedad de dominios, incluyendo la justicia penal, la banca, el acceso a la educación y la atención médica. En el ámbito de la justicia penal, algoritmos de predicción del riesgo de reincidencia han sido ampliamente criticados por su sesgo racial, ya que tienden a sobreestimar el riesgo de reincidencia de personas de color en comparación con las personas blancas. En la banca, los algoritmos de *scoring* crediticio enfrentan desafíos similares; la paridad demográfica se ha utilizado para mitigar el sesgo en la concesión de préstamos. Uno de los avances más significativos en el uso de métricas de *fairness* es el creciente uso de métodos de *fairness post-hoc*, que ajustan las decisiones algorítmicas después de que se ha entrenado el modelo. Estos métodos buscan corregir el sesgo sin alterar la estructura interna del modelo, lo que ha demostrado ser particularmente útil en sectores en los que la equidad es fundamental, como la atención médica y la educación.

A pesar de los avances en la creación de métricas para medir la equidad en los modelos de IA, aún existen varias limitaciones significativas que deben ser abordadas. En primer lugar, no hay una única métrica que garantice la equidad de manera universal. Cada métrica de *fairness* responde a un contexto específico y puede ser más o menos adecuada dependiendo del tipo de proyecto, la naturaleza de los datos y las consecuencias de los errores algorítmicos. Este es uno de los mayores desafíos a los que se enfrentan los investigadores de IA ética y los desarrolladores de estos modelos. Las métricas de distintos tipos, como la paridad demográfica, la paridad de resultados, el impacto dispar y la justicia individual, aunque útiles, no pueden cubrir todos los posibles escenarios de sesgo y discriminación (Selbst *et al*. 2019). De hecho, algunas de estas métricas incluso pueden entrar en conflicto entre sí: "[e]s matemáticamente imposible satisfacer simultáneamente todas las métricas de equidad, excepto en situaciones muy específicas" (Binns 2018, 2).

Por ejemplo, optimizar la paridad demográfica puede resultar en una disminución de la paridad de resultados del modelo, especialmente si los grupos demográficos no están representados de manera equitativa en los datos de entrenamiento. Esta situación puede llevar a una paradoja: para alcanzar la equidad entre grupos, el modelo debe renunciar a cierta precisión general. En un contexto donde los errores tienen consecuencias serias, como en la atención médica o la justicia penal, esta pérdida de precisión puede ser inaceptable. Asimismo, optimizar la justicia individual puede entrar en conflicto con la paridad de resultados, ya que garantizar que todos los individuos con características similares reciban el mismo tratamiento no implica necesariamente que las tasas de falsos positivos y falsos negativos sean equitativas entre grupos demográficos.

Otra limitación importante es la dificultad técnica para implementar y comparar estas métricas en la práctica. Los desarrolladores a menudo deben hacer concesiones entre diferentes métricas y no existe un consenso claro sobre cuáles priorizar en cada caso (Binns 2018). Esta falta de estandarización dificulta la adopción generalizada de prácticas justas en la industria y plantea desafíos regulatorios. Además, las métricas de *fairness* no siempre son interpretables por los usuarios finales o los responsables de la toma de decisiones, lo que complica su integración en procesos de toma de decisiones transparentes

y responsables. En efecto, muchas veces las métricas van dirigidas a un público experto, cuestión que solo permite cumplir con la accesibilidad, pero no con la explicabilidad que se deriva de los estándares asociados a la transparencia algorítmica (Lapostol Piderit, Garrido Iglesias y Hermosilla Cornejo 2023).

Un aspecto crucial que también limita la efectividad de las métricas es su dependencia de los datos de entrenamiento. Si los datos de entrada están sesgados, las métricas pueden no ser suficientes para corregir los sesgos inherentes a los modelos, si no se aplican las técnicas de mitigación necesarias. Además, lograr corregir estos sesgos parte de la premisa de que los sesgos posibles ya han sido identificados o anticipados en los datos de entrenamiento, ya que los sesgos desconocidos no se pueden medir (Buyl y De Bie 2024). Este desafío es particularmente evidente en contextos donde los datos históricos reflejan desigualdades sociales profundas, lo que puede perpetuar la discriminación a través de los algoritmos. En otras palabras, un diseño adecuado de las métricas supone una preocupación por los datos de entrenamiento, validación y prueba, que incluso obligue a la intervención de esos conjuntos de datos para evitar sesgos muestrales, de etiquetado o de validación; esto se puede lograr con datos sintéticos y usarse para generar datos adicionales de grupos subrepresentados en el conjunto de datos original. Tal medida ayudaría a equilibrar la distribución de datos entre los grupos y a reducir el sesgo del modelo (Ungern-Sternberg 2022).

Por ejemplo, si un algoritmo de contratación en el Estado se entrena con un conjunto de datos en el que las mujeres están infrarrepresentadas en puestos de "alta dirección pública", se pueden generar datos sintéticos de mujeres con perfiles de liderazgo para mejorar la equidad del algoritmo. Además, es posible elaborar datos sintéticos que simulen escenarios hipotéticos donde se modifiquen las características sensibles de los individuos, manteniendo constantes otros factores relevantes. Esto permite evaluar cómo el algoritmo se comporta en diferentes escenarios, detectar posibles sesgos y evitar resultados discriminatorios (Wachter, Mittelstadt y Russell 2018). Por ejemplo, se puede crear un conjunto de datos sintético en el que el género de los solicitantes de un préstamo se invierta para analizar si el algoritmo otorga préstamos de forma equitativa a hombres y mujeres con el mismo perfil financiero.

Finalmente, también hay que tener en cuenta las limitaciones legales y regulatorias. A medida que los Gobiernos y organizaciones internacionales desarrollan marcos regulatorios para la IA, las métricas de *fairness* deben alinearse con los requisitos legales, como las normativas sobre derechos humanos y no discriminación. Sin embargo, muchas veces estos marcos no están claramente definidos o armonizados entre distintas jurisdicciones, lo que produce incertidumbre sobre qué métrica utilizar y en qué medida los resultados algorítmicos cumplen con las regulaciones. Además, la legislación antidiscriminatoria, especialmente en el contexto europeo, se caracteriza por un enfoque contextual, en el que la evaluación de la discriminación se realiza caso por caso considerando las particularidades del contexto social y legal. Las métricas estadísticas de equidad, por su naturaleza cuantitativa, pueden tener dificultades para capturar esta complejidad contextual, lo que complica su aplicación directa en la legislación (Wachter, Mittelstadt y Russell 2021).

Ahora bien, la pregunta más importante para definir una métrica de *fairness*, que nos permitirá evaluar o auditar algoritmos con base en estándares más o menos objetivos, dependerá de la finalidad, el problema o la pregunta que cada sistema de decisión automatizada esté llamado a resolver. Estas preguntas requieren ser abordadas con una perspectiva sociotécnica, que implica considerar no solo las bases o fundamentos normativos, sino las formas en que los seres humanos interactúan con sistemas automatizados de decisión.

Así, por ejemplo, los algoritmos públicos que se utilizan en el sistema penal, que implican adoptar decisiones determinantes para la libertad individual de las personas, están constreñidos por la vinculación de derechos fundamentales que adquieren un carácter prevalente. Las métricas que se centren únicamente en la equidad grupal, como la paridad demográfica, posiblemente resultarían en la violación de los derechos individuales. En un sistema de justicia penal que pretenda usar algoritmos públicos, aplicar una estricta paridad demográfica podría llevar a la detención de personas inocentes de un grupo sobrerrepresentado para igualar las tasas de detención de otro grupo, cuestión que sería absurda (Selbst *et al*. 2019). Entonces, de acuerdo con una interpretación estándar, las métricas a ser utilizadas por los algoritmos públicos en el sistema penal deberían minimizar los falsos positivos. Sin embargo, ello dependerá de la finalidad del algoritmo: si lo que está en juego es la evaluación del riesgo de una víctima de violencia de género de sufrir un nuevo atentado contra su vida o integridad física o psíquica, entonces quizás la métrica a ser utilizada motivará la disminución de los falsos negativos, es decir, de aquellos casos que fueron considerados como de bajo riesgo y que motivaron a la Fiscalía o al Ministerio Público a no solicitar una medida cautelar más gravosa contra el imputado; por otra parte, si el mismo algoritmo se analiza ahora desde la perspectiva del imputado, la métrica preferida debería optar por la disminución de los falsos positivos, es decir, aquellos individuos que han sido catalogados como un peligro inminente para la víctima y que no lo son en realidad.

En este último tipo de casos, es fundamental evitar que un algoritmo clasifique erróneamente a un individuo como un riesgo cuando no lo es, especialmente cuando esta clasificación puede tener graves consecuencias para su libertad. Esta complejidad y las tensiones inevitables entre diferentes objetivos y métricas de sesgos reflejan por qué el problema de diseñar algoritmos justos y socialmente aceptables en contextos sensibles, como el sistema penal, es considerado por muchos como extremadamente desafiante. La imposibilidad de satisfacer simultáneamente todos los criterios de equidad estadística obliga a tomar decisiones que implican compromisos, lo que subraya la necesidad de un análisis multidimensional que integre aspectos técnicos, éticos y legales.

En este escenario, para evaluar la equidad de un algoritmo, a veces es crucial analizar si la proporción entre las tasas de falsos positivos y falsos negativos es la misma para los distintos grupos a los que el algoritmo aplica. A esta métrica se le conoce como *error ratio parity* (ERP). Aunque esta métrica no garantiza por sí sola que un algoritmo sea justo o injusto, permite detectar indicios relevantes para la evaluación de la justicia algorítmica. Los algoritmos, como cualquier herramienta que pretende modelar la realidad para adoptar decisiones de acuerdo con ciertos parámetros o finalidades propias, son imperfectos y pueden cometer errores en dos direcciones: identificar erróneamente a alguien como perteneciente a una categoría (falso positivo) o no identificar a alguien que sí pertenece (falso negativo). Lo determinante aquí, según Deborah Hellman (2020), es que el diseño del algoritmo debe equilibrar los costos de ambos errores según el contexto. Por ejemplo, en seguridad aérea, los falsos negativos (no identificar a un terrorista) son más costosos que los falsos positivos (detener a alguien inocente). En cambio, en el sistema penal, los falsos positivos (condenar a un inocente) suelen ser considerados más graves. En este último caso, como saben los abogados, se adopta la fórmula de Blackstone (también conocida como *ratio* de Blackstone), principio que establece que "es mejor que diez personas culpables escapen a que un inocente sufra". Cuando un algoritmo aplica reglas distintas para grupos diferentes (ya sea explícita o implícitamente a través de la ERP), se genera un tratamiento dispar. Esto puede ocurrir, por ejemplo, si se valora de forma desigual los costos de los errores para distintos grupos.

Un caso emblemático es el uso del algoritmo Compas que, a pesar de tener buenos indicadores en materia de precisión predictiva de reincidencia, mostró disparidades en sus tasas de error para personas negras y blancas, lo que sugiere un desequilibrio en cómo

se evaluaban los riesgos según el grupo racial (Larson *et al*. 2016). El problema, en este caso, consiste en cómo analizar el impacto de los errores del algoritmo. De acuerdo con la investigación de ProPublica, las personas negras tienen casi dos veces mayor probabilidad que las personas blancas de ser etiquetadas como de mayor riesgo de reincidencia delictiva, aunque no sea el caso; por otra parte, el algoritmo Compas comete el error opuesto entre las personas blancas: tienen mucha más probabilidad que las negras de ser etiquetadas como de menor riesgo, pero aun así terminan reincidiendo (Larson *et al*. 2016). Una propuesta para prevenir este tipo de consecuencia es intervenir el algoritmo de modo que incorpore un nivel de confianza desigual a favor de las personas negras. Sin embargo, y tal como notamos antes, uno de los problemas es la prohibición general de recolectar datos sensibles y procesarlos con el objeto de determinar efectos jurídicos.

En resumen, garantizar la equidad algorítmica implica buscar un equilibrio comparable en la relación entre falsos positivos y falsos negativos para todos los grupos relevantes. La falta de esta paridad puede reflejar y perpetuar desigualdades sistémicas.

Sin embargo, en otros contextos la cuestión no depende tanto de la *ratio* de los errores, sino de la *ratio* entre grupos. Al elegir métricas para asegurar la equidad en un algoritmo de contratación, es crucial priorizar un enfoque que proteja los derechos individuales y promueva la igualdad de oportunidades, sin dejar de lado la eficiencia y la eficacia en la selección de personal. En estos contextos, podemos democráticamente decidir que lo que importa no es tanto el costo del error algorítmico, sino el resultado final de la decisión, evaluado según ciertos parámetros que consideremos que articulan de manera adecuada todos los principios que señalamos antes: la igualdad de oportunidades, los derechos fundamentales y la eficacia en el reclutamiento y la selección de personal talentoso. Democráticamente, es posible decidir que los empleadores, que controlan el acceso a posiciones relevantes tanto para el reconocimiento personal como para la redistribución de ingresos, deberían tolerar ciertos umbrales cuantitativos evaluados según el resultado de sus procedimientos de contratación. Además, si estamos hablando del derecho a acceder en igualdad de oportunidades a empleos públicos, este tipo de decisiones podría tener un mayor sustento: el Estado no solo necesita reclutar el mejor talento posible, sino también usar el empleo público como una forma de reparar desventajas sociales, lo que podría tener efectos positivos en otros ámbitos (Harrisr y Foster 2010). Aún más, este tipo de decisiones, que llevan a la adopción de una métrica de paridad demográfica o de impacto dispar, podría justificarse mediante teorías consecuencialistas que den cuenta del derecho antidiscriminación y de sofisticados esquemas de prohibiciones de discriminación indirecta basados en la idea de que este tipo de regímenes están legitimados porque maximizan el bienestar de quienes están en peor situación. En tal caso, no sería tanto una cuestión de igualdad, sino un modo para maximizar el bienestar de quienes tienen menos acceso a los medios para llevar una vida digna (Lippert-Rasmussen 2013).

Un caso emblemático en este sentido es la métrica adoptada por la regla del 80% de la Comisión para la Igualdad de Oportunidades en el Empleo (EEOC, por sus siglas en inglés), que se basa en la tasa de selección de los postulantes, desagregados según las categorías protegidas por el derecho antidiscriminación federal aplicable en materia laboral (US Equal Employment Opportunity Commission 2022). Esta regla establece que la tasa de selección para un grupo protegido no debe ser inferior al 80% de la tasa de selección del grupo con la tasa más alta. Por ejemplo, si la tasa de selección para hombres en un puesto de trabajo es del 60%, la tasa de selección para mujeres en el mismo puesto no debería ser inferior al 48% (60 × 0,8 = 48). Si la tasa de selección para mujeres es inferior al 48%, se presume que existe una disparidad de impacto y el empleador debe justificar el uso de la práctica de contratación que está produciendo la disparidad. La métrica específica utilizada en la regla del 80% es la *ratio* entre las tasas de selección de los dos grupos. Se calcula dividiendo la tasa de selección del grupo protegido por la tasa de selección del grupo con la tasa más alta. Si el resultado es inferior a 0,8, se considera que existe una disparidad de impacto.

En otros casos, además, podríamos utilizar una métrica de precisión predictiva, que es útil en contextos donde el objetivo principal es la exactitud, eficacia o eficiencia de las predicciones, sin que las implicancias para la equidad o la justicia sean significativas. En otras palabras, se trataría de casos en los que lo que importa es el porcentaje de predicciones acertadas con respecto al total. Aquí, lo relevante es que el sistema automatizado de decisión trate a todas las personas de la manera más imparcial posible, optimizando un interés público prevalente que justifique la automatización de este proceso.

En ocasiones, la precisión predictiva puede ser el factor más importante, por ejemplo, en la detección de fraudes financieros o en el diagnóstico médico. En estas situaciones, la finalidad de tales sistemas de decisión parece ser el diagnóstico de la mayor cantidad de casos con enfermedades, lo que justificaría su uso, sin perjuicio de adoptar las medidas para abordar los falsos negativos. Otra circunstancia relevante en la que la precisión predictiva parece estar justificada es cuando existe un interés público prevalente, como en los algoritmos de fiscalización del cumplimiento de normativa tributaria o laboral. En efecto, siempre hay derechos fundamentales que los administrados podrían invocar, sobre todo en casos que dependen del denominado "debido proceso administrativo", pero no existe un derecho fundamental a oponerse a una fiscalización, especialmente si hay indicios fundados de un riesgo de incumplimiento de la normativa tributaria.

Quizás, en casos excepcionales, las potestades de fiscalización puedan utilizarse con fines discriminatorios, pero siempre queda la posibilidad de ejercer el derecho a la tutela judicial efectiva que permita la revisión judicial de aquella potestad. En otras palabras, la racionalización de los recursos involucrados en el ejercicio de potestades fiscalizadoras, con la finalidad última de obtener la mayor recaudación tributaria posible, justificaría minimizar la importancia de un análisis de las tasas de error o de disparidad entre grupos. Ello sugeriría que no estamos ante una métrica de equidad estadística, sino ante un algoritmo cuyo objetivo primario justifica minimizar la importancia de métricas distintas. Desde esta perspectiva, cualquier esfuerzo por tratar de mejorar la equidad estadística supondría necesariamente disminuir la precisión del algoritmo, que ha sido una de las premisas tradicionales en la ciencia de datos para tratar de balancear los intereses en cuestión (Corbett-Davies *et al*. 2017).

En un caso reciente ocurrido ante el órgano que supervisa el cumplimiento de la normativa chilena de transparencia y acceso a la información, un contribuyente solicitó conocer la "forma en que el análisis de datos masivos, *big data*, minería de datos y procesamientos de datos" influye en las decisiones de la autoridad administrativa tributaria, lo referido "a la práctica de fiscalizaciones a los contribuyentes", a "los parámetros algorítmicos utilizados para determinar que un contribuyente es más o menos susceptible a la comisión de una falta", así como también "los criterios utilizados para iniciar un proceso de fiscalización cuando se realice en virtud del análisis de datos masivos" (Consejo para la Transparencia 2023, considerando cuarto). Finalmente, sabiendo del amparo que interpuso el contribuyente ante la negativa del mencionado servicio, el Consejo para la Transparencia resolvió que el marco normativo que aloja el denominado "algoritmo de fiscalización centralizada" permite mantener cierta reserva, siempre y cuando muestre que este sistema automatizado de decisión observa niveles aceptables de predicción que permitan al servicio focalizar sus recursos y optimizar su estrategia de fiscalización. El propio Servicio de Impuestos Internos (SII) señaló que "la estrategia de prevención y control del cumplimiento tributario que ha desarrollado el SII apunta a generar acciones de tratamiento proporcionales a los niveles de incumplimiento, tamaño y riesgo, a partir del análisis del comportamiento de las y los contribuyentes, a través de modelos de analítica avanzada". En tal sentido, consideró el CPLT, la reserva del SII se encuentra plenamente justificada, ya que se trataría de

> información sensible [...] cuya divulgación conlleva la posibilidad cierta de afectación del debido cumplimiento de la función fiscalizadora [...], en atención a que la información, antecedentes y actuaciones realizadas por el organismo en virtud de los parámetros algorítmicos referidos, contiene los ámbitos, métodos de trabajo y mecanismos específicos de fiscalización, de prevención de actuaciones irregulares y de pesquisa temprana de situaciones de riesgo de incumplimientos e irregularidades tributarias; de procesos internos de tratamiento de datos de los contribuyentes; de las medidas de control que resultan necesarias para validar eventuales modificaciones y/o actualizaciones de la información, todo lo cual en caso de ser divulgado configura un riesgo de daño cierto, probable y específico a la tarea de fiscalización que debe efectuar el SII en cumplimiento de un mandato legal, con la consecuente afectación del debido cumplimiento de las funciones del órgano requerido y de los intereses económicos del Estado, por las implicancias que este daño generaría en la recaudación tributaria.

De este modo, la divulgación de la información solicitada por el contribuyente vulneraría "la eficacia de determinados procesos de fiscalización y la función fiscalizadora", y podría "ser utilizada por terceros para burlar las labores de control y fiscalización, y, en definitiva, para incumplir la normativa tributaria". A su vez, el Consejo señaló que ello tendría como consecuencia una afectación al "interés nacional", específicamente a "los intereses económicos del país atendida la naturaleza de las funciones que desarrolla" (Consejo para la Transparencia 2023, considerando cuarto).

En la tabla 1 proponemos diferentes criterios para seleccionar métricas de *fairness* en contextos de uso algorítmico por parte del Estado:

Tabla 1. Criterios de selección de métricas de *fairness* por parte del Estado

Contexto	Finalidad principal	Criterio de prioridad	Métrica sugerida	Justificación
Sistema procesal penal (decisiones judiciales con respecto a imputados)	Proteger la libertad individual de las personas sometidas a procesos penales	Minimizar falsos positivos	*Error ratio parity* y paridad de tasa de falsos positivos	En el caso de imputados, evitar clasificaciones erróneas que lleven a restricciones injustificadas de la libertad y detectar si un grupo experimenta un trato dispar en términos de errores algorítmicos (falsos positivos/negativos).
Sistema procesal penal (protección de víctimas)	Prevenir riesgos de violencia y proteger la vida y la integridad	Minimizar falsos negativos	Tasa de falsos negativos y tasa de falsa omisión	Reducir casos en los que se subestime el riesgo de violencia y no se adopten medidas cautelares necesarias.
Algoritmos de contratación pública	Promover igualdad de oportunidades	Equidad en las tasas de selección	Impacto dispar y paridad demográfica	Garantizar que grupos protegidos no sean discriminados indirectamente durante el proceso de selección.
Algoritmos de diagnóstico médico en hospitales públicos	Maximizar detección de casos positivos	Maximizar precisión predictiva	Precisión predictiva y paridad de tasa de falsos positivos	Identificar la mayor cantidad de diagnósticos correctos sin comprometer la seguridad del paciente.

Contexto	Finalidad principal	Criterio de prioridad	Métrica sugerida	Justificación
Seguridad aérea	Prevenir actos terroristas o atentados a la seguridad	Minimizar falsos negativos	Tasa de falsos negativos y tasa de falsa omisión	Asegurar que ningún individuo de alto riesgo pase inadvertido por el sistema de detección.
Fiscalización tributaria/laboral	Optimizar recursos y garantizar el cumplimiento normativo	Maximizar la eficacia global (precisión)	Precisión predictiva	Lograr un uso eficiente de recursos al focalizar fiscalizaciones basadas en modelos de riesgo confiables.

Fuente: elaboración propia.

La tabla introduce una serie de pautas normativas al sugerir qué métricas de equidad deberían priorizarse en función del contexto de uso del algoritmo. No obstante, es importante explicitar que esta propuesta no pretende deducir directamente los criterios de equidad a partir de los objetivos primarios del sistema automatizado. Más bien, la intención es identificar qué métrica puede resultar más adecuada para evaluar la equidad en relación con los riesgos que enfrenta cada grupo afectado por el algoritmo, con el contexto en el que se implementa el sistema, y con los principios constitucionales o derechos fundamentales comprometidos en cada dominio de uso.

Así, por ejemplo, si en el sistema penal se propone priorizar métricas que minimicen falsos positivos en relación con personas imputadas, no es porque este sea el único objetivo del sistema, sino porque las consecuencias jurídicas de una clasificación errónea en este contexto —la privación de la libertad de una persona inocente— tienen un peso especialmente grave desde una perspectiva de derechos fundamentales. Es cierto que esta priorización puede entrar en tensión con otros objetivos, como la paridad en la tasa de predicciones o la eficacia general del sistema. No obstante, estas tensiones no pueden resolverse desde un único criterio técnico: requieren deliberación pública y principios normativos que orienten la toma de decisiones sobre qué errores son socialmente tolerables o aceptables.

El caso del ASR es emblemático en este sentido, pues obliga a considerar los contextos y situaciones en que este algoritmo se utiliza con el objetivo de identificar de manera eficiente a las mujeres víctimas de violencia de género por delitos de violencia intrafamiliar. En estos casos, las medidas de protección pueden ir directamente dirigidas a la víctima y no involucrar una afectación de los derechos del imputado. Sin embargo, tanto en el diseño como en la implementación práctica del ASR, las consideraciones sobre los derechos del imputado deben entrar a jugar un rol fundamental. De este modo, la discusión sobre qué aspectos automatizar, cuánto ahondar en la necesidad de explicar las variables de ponderación que explicarían un alto riesgo, o cuándo sujetar ciertas decisiones a un doble escrutinio o a una intervención humana son todos aspectos que influyen de manera determinante en la pregunta por qué métrica de equidad estadística podríamos utilizar para ayudarnos a evaluar la justicia del algoritmo público en un caso concreto.

En efecto, los contextos no son compartimentos estancos y, en muchos casos, las fronteras entre categorías como *protección de víctimas* o *procesamiento de imputados* son más difusas de lo que la tipología sugiere. El sistema procesal penal, como señalamos antes, es un complejo engranaje en el que todas las partes, sus vínculos e interacciones están conectados. En este sentido, la tabla debe leerse como un ejercicio analítico inicial y no como una clasificación cerrada. De hecho, sería perfectamente razonable sostener que distintas métricas pueden ser necesarias dentro de un mismo dominio —por ejemplo, dolencias médicas específicas requerirían diferentes criterios de equidad, como ocurre

con el diagnóstico de enfermedades graves versus enfermedades leves—, cuestión que supone un ejercicio bastante más complejo que la selección de métricas de equidad estadísticas, que constituyen un mero complemento a la difícil pregunta por la justicia de las decisiones que adopta un órgano público. Así, podemos superar la barrera de la imposibilidad matemática de distintas métricas de equidad estadística en un mismo modelo de decisión, y quizás proponer la posibilidad de complementarlas en el marco de un sistema que, como el procesal penal, supone equilibrar demasiados intereses y valores en juego dependiendo del contexto.

La selección de la métrica adecuada, por tanto, no puede reducirse a una decisión técnica; requiere una perspectiva sociotécnica que considere las implicancias normativas y las dinámicas humanas en la interacción con los sistemas automatizados, buscando un equilibrio entre los costos de los errores, la equidad entre grupos y la protección de derechos individuales en cada contexto.

Conclusión

La equidad algorítmica representa un desafío complejo debido a la variedad de métricas disponibles y a la necesidad de contextualizar su aplicación según los valores éticos, sociales y jurídicos en juego. Diversas investigaciones han revelado que no existe una métrica universalmente adecuada, ya que cada una aborda aspectos específicos de un problema. Por ejemplo, la paridad demográfica es útil para garantizar igualdad de oportunidades en procesos de selección, pero puede resultar inapropiada en contextos en los que la precisión individual es crítica, como en diagnósticos médicos o predicción de riesgos. En cambio, métricas como el ERP, que evalúan la distribución de errores entre grupos, son esenciales en ámbitos como el sistema procesal penal, en el que minimizar falsos positivos o negativos de acuerdo con el impacto que sufren ciertos grupos desaventajados tiene consecuencias directas e intensas sobre un cúmulo de derechos fundamentales.

A través del análisis de distintos algoritmos públicos incluidos en un repositorio desarrollado por la Universidad Adolfo Ibáñez de Chile, mostramos cómo los sesgos algorítmicos pueden surgir en diversas etapas del ciclo de vida de los modelos: desde la recolección de datos hasta su implementación. Sesgos sociales, de representación, de etiquetado o de medición pueden introducir desigualdades en los resultados, incluso si los modelos se ajustan a criterios matemáticos de equidad. Además, las métricas no solo deben garantizar la equidad estadística entre grupos, sino también respetar los principios del derecho a la igualdad y a la no discriminación, evitando decisiones que perpetúen injusticias estructurales.

A diferencia de cierta literatura especializada, nuestra investigación no busca establecer un protocolo ético definitivo ni un árbol de decisión concreto para seleccionar métricas de equidad (véase, por ejemplo, Ruf y Detyniecki [2021]), sino identificar las preguntas que deben ser enfrentadas de manera crítica al evaluar su correspondencia con estándares jurídicos. En este sentido, es clave considerar lo planteado por Buijsman (2024), quien advierte que, dada la imposibilidad matemática de optimizar simultáneamente todas las métricas de equidad y la tensión estructural entre precisión y equidad, se requiere una teoría sustantiva que guíe nuestras decisiones. Su propuesta, basada en la noción rawlsiana de justicia como equidad, subraya la importancia de priorizar aquellas métricas que más impactan el bienestar de los grupos más vulnerables; se cierra así parte de la brecha entre los enfoques filosóficos de justicia distributiva y el uso técnico de métricas de equidad. A diferencia de Buijsman, el enfoque propuesto en este trabajo asume una perspectiva jurídica que, si bien dialoga con nociones filosóficas como la de justicia distributiva, no se encuentra necesariamente anclada en ellas.

En contextos institucionales diversos, los sistemas de IA deben ser evaluados conforme a estándares normativos específicos, como el principio de igualdad ante la ley, el derecho a la no discriminación, la razonabilidad de las decisiones o la proporcionalidad en la afectación de derechos. Estos estándares varían según el ámbito de aplicación —por ejemplo, justicia penal, políticas sociales o salud— y exigen, más que una teoría moral general, una interpretación contextualizada de los marcos normativos aplicables. Desde esta óptica, la elección de una métrica de equidad no puede desvincularse del análisis jurídico de las obligaciones del Estado, del tipo de derechos en juego ni del tipo de justificación que exige el derecho administrativo o constitucional en cada caso. Así, este enfoque jurídico no pretende ofrecer respuestas únicas o universales, sino visibilizar que las decisiones sobre métricas de equidad son, en última instancia, decisiones jurídicas situadas.

De este modo, destacamos que el uso de métricas requiere un proceso de gobernanza interdisciplinaria. La colaboración entre científicos de datos, juristas y expertos en ciencias sociales resulta indispensable para regular, diseñar y auditar modelos algorítmicos de manera transparente y responsable. Garantizar la equidad algorítmica no implica perseguir una perfección técnica inalcanzable, sino lograr un balance adecuado entre precisión, justicia grupal e individual y protección de los derechos fundamentales en cada contexto específico.

Por último, queremos resaltar que el repositorio de Algoritmos Públicos desarrollado por la Universidad Adolfo Ibáñez de Chile representa una valiosa contribución para avanzar en la comprensión crítica de cómo se diseñan, implementan y evalúan sistemas automatizados en el sector público. Esta herramienta ha permitido vincular casos concretos con preguntas normativas relevantes sobre equidad, sesgo y derechos fundamentales, abriendo nuevas posibilidades para conectar el análisis técnico con marcos jurídicos específicos. En paralelo, la reciente promulgación de una nueva ley sobre protección de datos personales y el proyecto de ley de inteligencia artificial propuesto por el Gobierno chileno reflejan una creciente sensibilidad institucional ante los desafíos que plantea el uso de estas tecnologías. A partir de este escenario, surgen nuevas preguntas que podrían orientar futuras investigaciones: ¿cómo facilitar una aplicación efectiva de los principios de transparencia, explicabilidad y proporcionalidad en distintos contextos institucionales? ¿Qué tipo de capacidades estatales serían necesarias para auditar estos sistemas desde una perspectiva interdisciplinaria? ¿Y en qué medida es posible traducir estándares normativos en criterios operativos para seleccionar métricas de equidad adecuadas a cada caso? Estas preguntas, más que ofrecer certezas, invitan a profundizar en el diálogo entre lo técnico y lo jurídico, y a seguir explorando cómo lograr una implementación más justa y socialmente contextualizada de la inteligencia artificial en el ámbito público.

Referencias

1. Bekkum, Marvin van. 2024. "Using Sensitive Data to Debias AI Systems: Article 10 (5) of the EU AI Act". SSRN. https://dx.doi.org/10.2139/ssrn.4992036
2. Binns, Reuben. 2018. "Fairness in Machine Learning: Lessons from Political Philosophy". En *Proceedings of Machine Learning Research* 81: 149-159. Conference on Fairness, Accountability, and Transparency, 23 al 24 de febrero, Nueva York, Estados Unidos. https://proceedings.mlr.press/v81/binns18a/binns18a.pdf
3. Buijsman, Stefan. 2024. "Navigating Fairness Measures and Trade-Offs". *AI Ethics* 4: 1323-1334. https://doi.org/10.1007/s43681-023-00318-0
4. Buyl, Maarten y Tijl De Bie. 2024. "Inherent Limitations of AI Fairness". *Communications of the ACM* 67 (2): 48-55. https://doi.org/10.1145/3624700
5. Caton, Simon y Christian Haas. 2024. "Fairness in Machine Learning: A Survey". *ACM Computing Surveys* 56 (7): 1-38. https://doi.org/10.1145/3616865
6. Chacón, Andrés, Eduardo E. Kausel y Tamara Reyes. 2022. "A Longitudinal Approach for Understanding Algorithm Use". *Journal of Behavioral Decision Making* 35 (4): e2275. https://doi.org/10.1002/bdm.2275

7. Coddou McManus, Alberto, Isabel Ruiz-Esquide, Dominique Bergeret, Camila Loyola y Valentina Sánchez. En prensa. "IA y violencia contra las mujeres: la automatización de la evaluación del riesgo". *Política Criminal*.
8. Consejo para la Transparencia. 2023. Decisión amparo ROL C12867-23, 29 de noviembre de 2023. Entidad pública: Servicio de Impuestos Internos. Requirente: Mauricio Álvarez Vega. https://jurisprudencia.cplt.cl/Paginas/ResultadoBusqueda.aspx?data=86C88158961C
9. Contreras, Pablo. 2024. "Convergencia internacional y caminos propios: regulación de la inteligencia artificial en América Latina". *Actualidad Jurídica Iberoamericana* 21: 468-491. https://revista-aji.com/convergencia-internacional-y-caminos-propios-regulacion-de-la-inteligencia-artificial-en-america-latina/
10. Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel y Aziz Hug. 2017. "Algorithmic Decision Making and the Cost of Fairness". En *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Minin*, 797-806. 13 al 17 de agosto, Halifax, Canadá. https://doi.org/10.1145/3097983.3098095
11. Giffen, Benjamin van, Dennis Herhausen y Timo Fahse. 2022. "Overcoming the Pitfalls and Perils of Algorithms: A Classification of Machine Learning Biases and Mitigation Methods". *Journal of Business Research* 144: 93-106. https://doi.org/10.1016/j.jbusres.2022.01.076
12. Green, Ben. 2022. "Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness". *Philosophy & Technology* 35 (90): en línea. https://doi.org/10.1007/s13347-022-00584-6
13. Harrisr, Lynette y Carley Foster. 2010. "Aligning Talent Management with Approaches to Equality and Diversity: Challenges for UK Public Sector Managers". *Equality, Diversity and Inclusion: An International Journal* 29 (5): 422-435. https://doi.org/10.1108/02610151011052753
14. Hellman, Deborah. 2020. "Measuring Algorithmic Fairness". *Virginia Law Review* 106 (4): 811-866. https://virginialawreview.org/wp-content/uploads/2020/06/Hellman_Book.pdf
15. Hermosilla, María Paz y Mariana Germán. 2024. "Implementación responsable de algoritmos e inteligencia artificial en el sector público de Chile". *Revista Chilena de la Administración del Estado* 11: 101-122. https://doi.org/10.57211/revista.v11i11.185
16. Lapostol Piderit, José Pablo, Romina Garrido Iglesias y María Paz Hermosilla Cornejo. 2023. "Algorithmic Transparency from the South: Examining the State of Algorithmic Transparency in Chile's Public Administration Algorithms". En *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 227-235. 12 al 15 de junio, Chicago, Estados Unidos. https://doi.org/10.1145/3593013.3593991
17. Larson, Jeff, Surya Mattu, Lauren Kirchner y Julia Angwin. 2016. "How We Analyzed the COMPAS Recidivism Algorithm". ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
18. Lippert-Rasmussen, Kasper. 2013. *Born Free and Equal?: A Philosophical Inquiry into the Nature of Discrimination.* Oxford: Oxford University Press.
19. Martínez Placencia, Victoria. 2023. "Funciones de las categorías de discriminación en el derecho laboral chileno". *Revista Chilena de Derecho* 50 (2): 1-31. https://doi.org/10.7764/R.502.1
20. Moreau, Sophia. 2010. "What Is Discrimination?". *Philosophy & Public Affairs* 38 (2): 143-179. https://doi.org/10.1111/j.1088-4963.2010.01181.x
21. Palma, Aníbal. 2024: "Data and Data Governance (Article 10)". En *The EU Regulation on Artificial Intelligence: A commentary*, 183-206. Roma: Wolters Kluwers.
22. Proyecto de ley del 7 de mayo de 2024. "Regula los sistemas de inteligencia artificial". Cámara de Diputadas y Diputados. https://www.camara.cl/legislacion/ProyectosDeLey/tramitacion.aspx?prmID=17429&prmBOLETIN=16821-19
23. RAE (Real Academia Española). 2025. *Diccionario de la lengua española*. 23.ª ed. https://dle.rae.es
24. Rawls, John. 2012. *Justicia como equidad: una reformulación.* Barcelona: Paidós.
25. Regulation (EU) 2024/1689 of the European Parliament and of the Council. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
26. Ruf, Boris y Marcin Detyniecki. 2021. "Towards the Right Kind of Fairness in AI". Cornell University. https://doi.org/10.48550/arXiv.2102.08453
27. Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian y Janet Vertesi. 2019. "Fairness and Abstraction in Sociotechnical Systems". En *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59-68. 29 al 31 de enero, Atlanta, Estados Unidos. https://doi.org/10.1145/3287560.3287598
28. Ungern-Sternberg, Andreas von. 2022. "Discriminatory AI and the Law: Legal Standards for Algorithmic Profiling". En *Responsible Artificial Intelligence*, editado por Silja Vöneky, Philipp Kellmeyer, Oliver Müller y Wolfram Burgard, 252-278. Cambridge: Cambridge University Press.
29. US Equal Employment Opportunity Commission. 2022. "The Americans with Disabilities Act and the Use of Software, Algorithms, and Artificial Intelligence to Assess Job Applicants and Employees". EEOC-NVTA-2022-2. https://perma.cc/LG25-D53T
30. Vandevelde, Kenneth J. 2010. "A Unified Theory of Fair and Equitable Treatment". *New York University Journal of International Law and Politics* 43 (1): 43-106. https://ssrn.com/abstract=2357642
31. Verma, Sahil y Julia Rubin. 2018. "Fairness Definitions Explained". En *FairWare '18: Proceedings of the International Workshop on Software Fairness*, 1-7. 29 de mayo, Gotemburgo, Suecia. https://doi.org/10.1145/3194770.3194776

32. Wachter, Sandra, Brent Mittelstadt y Chris Russell. 2018. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR". *Harvard Journal of Law & Technology* 31 (2): 841-887. https://doi.org/10.48550/arXiv.1711.00399

33. Wachter, Sandra, Brent Mittelstadt y Chris Russell. 2020. "Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law". *West Virginia Law Review* 123 (3): 3-51. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792772

34. Wachter, Sandra, Brent Mittelstadt y Chris Russell. 2021. "Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI". *Computer Law & Security Review* 41: 105567. https://doi.org/10.1016/j.clsr.2021.105567

Alberto Coddou Mc Manus

Doctor en Derecho por la Universidad de Londres, Reino Unido. Profesor asociado de College/Escuela de Gobierno de la Pontificia Universidad Católica de Chile. Sus líneas de investigación son la relación entre las tecnologías y los derechos fundamentales, principalmente a partir del análisis de algoritmos públicos y de aquellos implementados en el ámbito laboral. Últimas publicaciones: "Discriminación algorítmica en los procesos automatizados de reclutamiento y selección de personal" (en couatoría), *Revista Chilena de Derecho y Tecnología* 13: 186-219, 2024, https://doi.org/10.5354/0719-2584.2024.71312; y "Constitutional Change and Referendums in Chile and Ireland: Faraway, so Close" (en coautoría), *Politics and Governance* 13, 2025, https://doi.org/10.17645/pag.9197. https://orcid.org/0000-0003-2041-2304 | acoddou@uc.cl

Mariana Germán Ortiz

Magíster en Ciencias de Datos por la Universidad Adolfo Ibáñez, Chile. Investigadora de GobLab, laboratorio de innovación pública de la Escuela de Gobierno de la Universidad Adolfo Ibáñez, Chile. Sus líneas de investigación se centran en la inteligencia artificial ética y responsable, con especial foco en equidad, discriminación algorítmica y transparencia de los sistemas. Últimas publicaciones: "Applying the Ethics of AI: A Systematic Review of Tools for Developing and Assessing AI-Based Systems" (en coautoría), *Artificial Intelligence Review* 57 (110): 1-30, 2024, https://doi.org/10.1007/s10462-024-10740-3; e "Implementación responsable de algoritmos e inteligencia artificial en el sector público de Chile" (en coautoría), *Revista Chilena de la Administración del Estado* 11: 101-122, 2024, https://doi.org/10.57211/revista.v11i11.185. https://orcid.org/0009-0002-7360-1336 | mariana.german.o@uai.cl

Reinel Tabares Soto

Ph. D. en Ingeniería por la Universidad Autónoma de Manizales, Colombia. Profesor de planta e investigador del Departamento de Sistemas e Informática de la Universidad de Caldas, Colombia. Profesor del Departamento de Electrónica y Automatización de la Universidad Autónoma de Manizales, Colombia. Se desempeña como investigador y director alterno del proyecto "Algoritmos éticos y responsables" en la Universidad Adolfo Ibáñez, Chile. Sus líneas de investigación incluyen aprendizaje de máquina, minería de datos, computación de alto desempeño, bioinformática e inteligencia artificial responsable. Últimas publicaciones: "Building Better Forecasting Pipelines: A Generalizable Guide to Multi-Output Spatio-Temporal Forecasting" (en coautoría), *Expert Systems with Applications* 259 (1): 125384, 2025, https://doi.org/10.1016/j.eswa.2024.125384; "Predicting No-Shows at Outpatient Appointments in Internal Medicine Using Machine-Learning Models" (en coautoría), *PeerJ Computer Science* 9: e2762, 2025, https://peerj.com/articles/cs-2762. https://orcid.org/0000-0003-3639-4147 | reinel.tabares@ucaldas.edu.co