

Equidad, Rendición de Cuentas, y Transparencia en el Aprendizaje Automático para el caso la discriminación de género

**Introducción a los Modelos Computacionales. Grado en Ingeniería
Informática. Universidad de Córdoba. 2022-2023**

Javier Sánchez Monedero (Universidad de Córdoba)

Ana Valdivia García (Oxford University)

Objetivos

Parte I (Javier Sánchez)

- Introducción y motivación a FATE en inteligencia artificial
- Cuantificando y mitigando sesgos: [FairLearn](#)

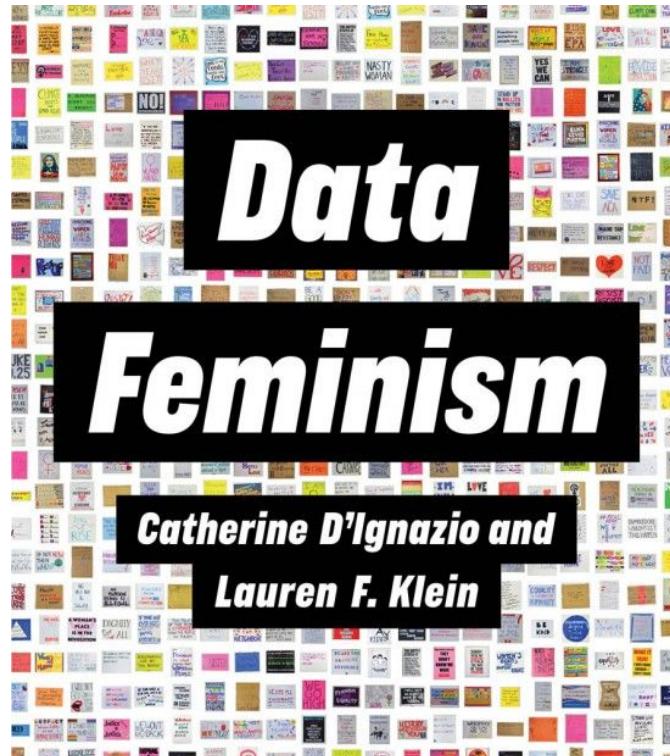
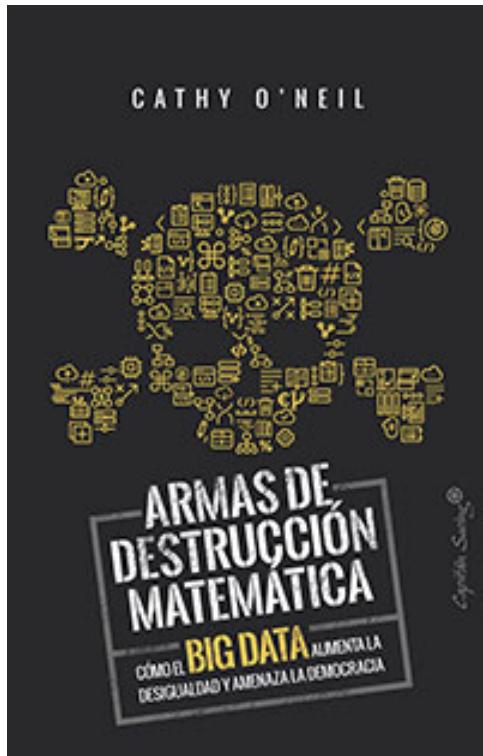
Parte II (Ana Valdivia)

- Analizando a un algoritmo interdisciplinariamente: [Judging the algorithm](#)



Introducción y motivación a FATE

¿Por dónde empezar? Libros



FAIRNESS AND MACHINE LEARNING

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

CONTENTS

PREFACE

ACKNOWLEDGMENTS

- 1 [INTRODUCTION](#) [PDF](#)

- 2 [WHEN IS AUTOMATED DECISION MAKING LEGITIMATE?](#) [PDF](#)

We explore what makes automated decision making a matter of normative concern, situated in bureaucratic decision making and its mechanical application of formalized rules.

- 3 [CLASSIFICATION](#) [PDF](#)

We introduce formal non-discrimination criteria in a decision-theoretic setting, establish their relationships, and illustrate their limitations.

- 4 [RELATIVE NOTIONS OF FAIRNESS](#) [PDF](#)

We introduce formal non-discrimination criteria in a decision-theoretic setting, establish their relationships, and illustrate their limitations.

- 5 [CAUSALITY](#) [PDF](#)

We dive into the rich technical repertoire of causal inference and how it helps articulate and address shortcomings of the classification paradigm, while raising new conceptual and normative questions.

¿Por dónde empezar? En vídeo

- Documental [Coded Bias](#)
- TED Talk de Joy Buolamwini
[How I'm fighting bias in algorithms](#)

CODED BIAS

THERE IS NO ALGORITHM FOR TRUTH



A SHALINI KANTAYYA FILM



FATE:

- **Fairness:**
imparcialidad/ecuanimidad
- **Accountability:** rendición de cuentas
- **Transparency:** transparencia
- **Ethics:** ética

facctconference.org

facctconference.org/network



Objetivos del seminario

Discriminación en **sistemas/modelos** que toman decisiones trascendentales

- Esto no considera otras formas de discriminación o injusticia
- Las cuestiones de discriminación/igualdad necesitan de otro tipo de intervenciones no técnicas (ver libros recomendados)

La discriminación no es un concepto general, depende:

- Dominio del problema
- Grupo social

La presentación de [Judging the algorithm](#) dará una visión más **interdisciplinar** de este problema.

Grupos protegidos

Clases protegidas (no en todos los contextos):

- EEUU: “raza”, color, sexo, religión, ciudadanía, embarazo...
- España: género, ley igualdad de trato, embarazo, ley igualdad de trato “raza”, embarazo...

La definición de grupos protegidos va más allá e incluye las categorías **no binarias** y la **interseccionalidad**

There's No Scientific Basis for Race—It's a Made-Up Label. National Geographic. 2018, March 12.



Ley integral igualdad de trato y no discriminación

Artículo 23 Ley 15/2022, de 12 de julio:

Artículo 23. *Inteligencia Artificial y mecanismos de toma de decisión automatizados.*

1. En el marco de la Estrategia Nacional de Inteligencia Artificial, de la Carta de Derechos Digitales y de las iniciativas europeas en torno a la Inteligencia Artificial, las administraciones públicas favorecerán la puesta en marcha de mecanismos para que los algoritmos involucrados en la toma de decisiones que se utilicen en las administraciones públicas tengan en cuenta criterios de minimización de sesgos, transparencia y rendición de cuentas, siempre que sea factible técnicamente. En estos mecanismos se incluirán su diseño y datos de entrenamiento, y abordarán su potencial impacto discriminatorio. Para lograr este fin, se promoverá la realización de evaluaciones de impacto que determinen el posible sesgo discriminatorio.

2. Las administraciones públicas, en el marco de sus competencias en el ámbito de los algoritmos involucrados en procesos de toma de decisiones, priorizarán la transparencia en el diseño y la implementación y la capacidad de interpretación de las decisiones adoptadas por los mismos.

3. Las administraciones públicas y las empresas promoverán el uso de una Inteligencia Artificial ética, confiable y respetuosa con los derechos fundamentales, siguiendo especialmente las recomendaciones de la Unión Europea en este sentido.

4. Se promoverá un sello de calidad de los algoritmos.

Las personas también tienen sesgos



Diferencias (O'Neil 2016):

- Sistematización
- Escala
- Nuevos grupos "digitales" discriminados

Casos: PNL + Visión Artificial

Algorithmic Bias in Grounded Setting



Casos: reconocimiento facial

Análisis interseccional del rendimiento en reconocimiento facial de Amazon Rekognition. La menor tasa de acierto se da para las mujeres de piel oscura.

August 2018 Accuracy on Facial Analysis Pilot Parliaments Benchmark

98.7% **68.6%** **100%** **92.9%**



**DARKER
MALES**



**DARKER
FEMALES**



**LIGHTER
MALES**

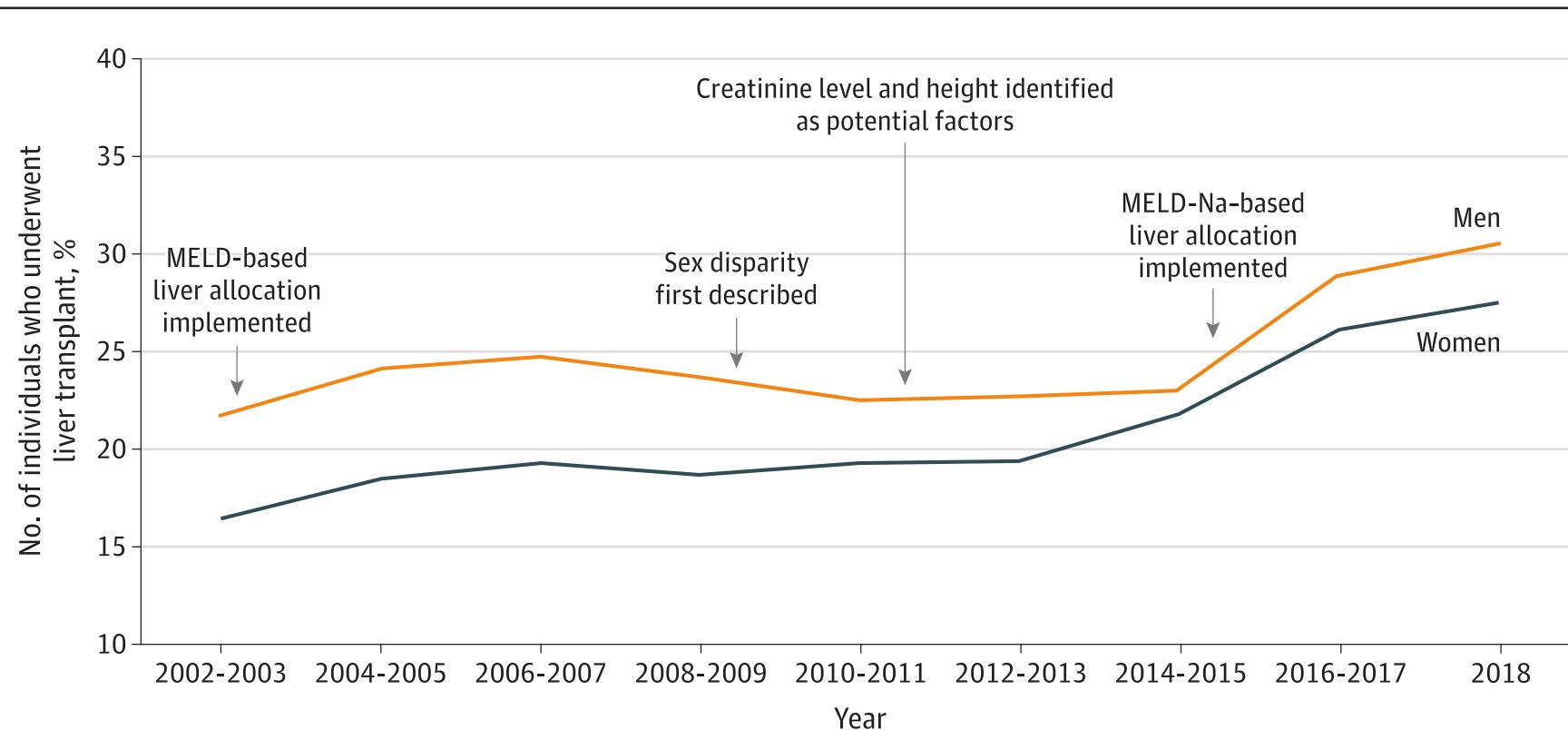


**LIGHTER
FEMALES**

Amazon Rekognition Performance on Gender Classification

Casos: biomedicina

Figure. Proportion of Women and Men Who Underwent Deceased Donor Liver Transplant and Timeline of Important Events in Liver Transplant



Verna, E. C., & Lai, J. C. (2020). Time for Action to Address the Persistent Sex-Based Disparity in Liver Transplant Access. *JAMA Surgery*, 155(7), 545–547. <https://doi.org/10.1001/jamasurg.2020.1126>

A1. ¿Cómo cuantificarías el sesgo en los problemas anteriores?

- **Reconocimiento facial:** el modelo tiene menos precisión identificando mujeres con piel oscura
- **Medicina:** el modelo subestima el riesgo de mujeres de morir en lista de espera
- **Procesamiento lenguaje natural:** el sistema reproduce estereotipos de género asociados a profesiones



REPORT
**AUTOMATING
SOCIETY**

2020



Bertelsmann Stiftung

EDICIÓN
EN ESPAÑOL

Inventarios de casos



ALGORITHM
WATCH

Automating Society Report 2020



OBSERVATORY OF
ALGORITHMS WITH
SOCIAL IMPACT

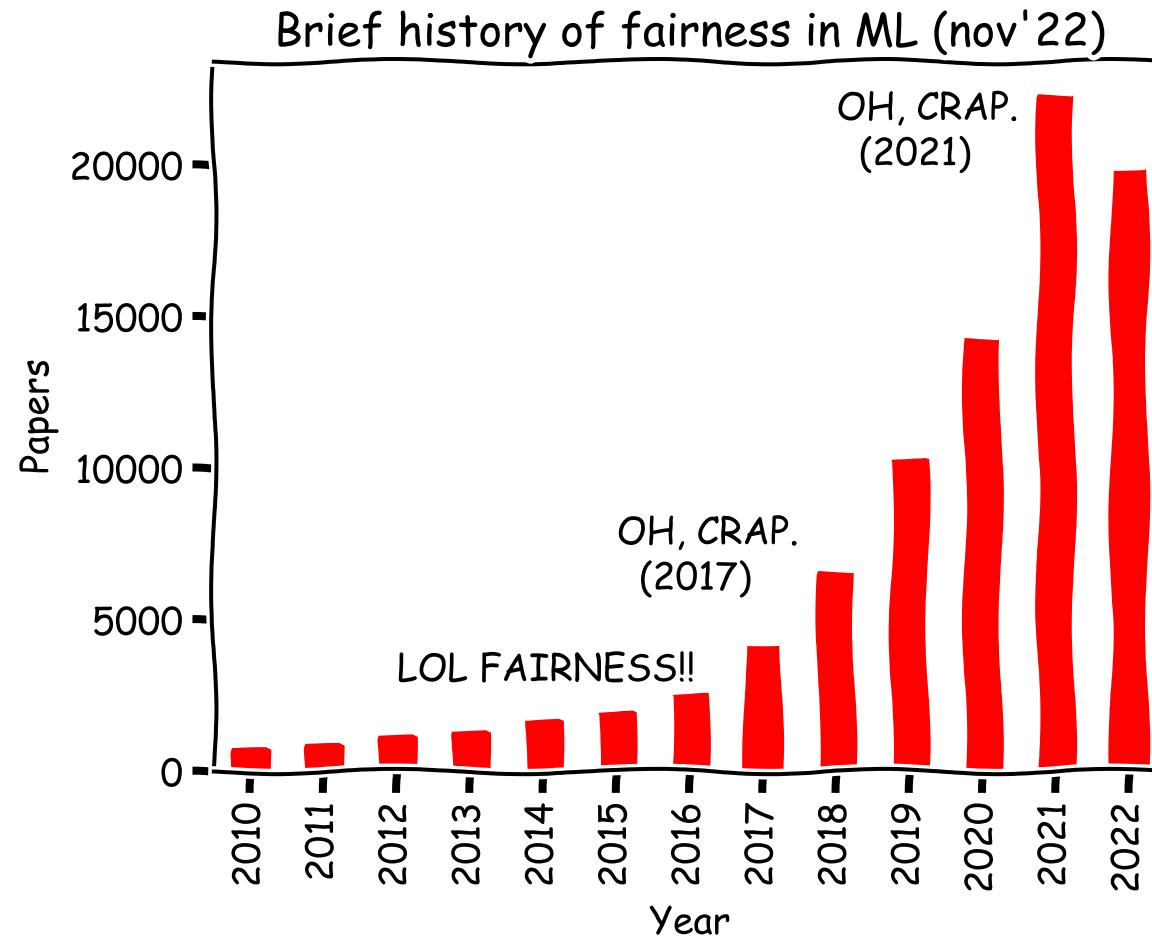
Observatory of Algorithms with Social
Impact

A2. Accede al informe de Algorithm Watch para ver si conoces estos sistemas. Igualmente entra en OASI y elige "Spain" para descubrir sistemas en uso.

Cuantificando y mitigando sesgos

¿Cómo medir y mitigar el sesgo?

Ecuanimidad sin hacer nada (*unawareness*)



Actualizada del NIPS 2017 Tutorial on Fairness in Machine Learning

Análisis exploratorio

- Comprobar distribución (prevalencia/prior) etiqueta de clase
- Comprobar distribución (prevalencia/prior) etiqueta de clase por grupos
- Comprobar:
 - Visual
 - Estadística descriptiva
 - Contraste de hipótesis

Un ejemplo excelente lo podeis ver en Straw, I., & Wu, H. (2022).

Straw, I., & Wu, H. (2022). Investigating for bias in healthcare algorithms: A sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health & Care Informatics*, 29(1), e100457. <https://doi.org/10.1136/bmjhci-2021-100457>

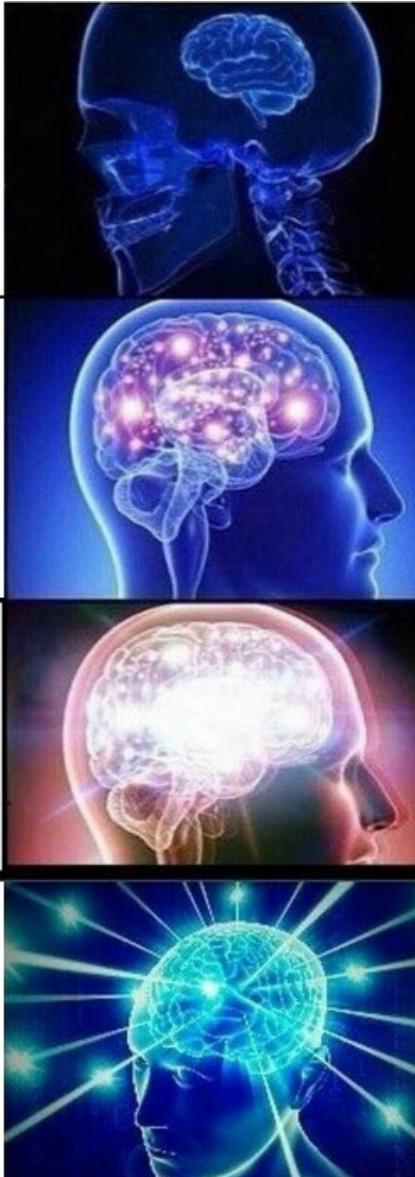
**MIRAR
LOS DATOS**

**MIRAR EL
HISTOGRAMA**

**ESTADÍSTICA
DESCRIPTIVA**

**CONTRASTE
DE HIPÓTESIS**

imgflip.com



El "zoo" de las métricas de ecuanimidad

	notion	use of Y	condition
group fairness	Demographic Parity	-	equal acceptance rate across groups
	Conditional Demographic Parity	-*	equal acceptance rate across groups in any strata
	Equal Accuracy	✓	equal accuracy across groups
	error parity	Equality of Odds Predictive Parity	✓ ✓ equal FPR and FNR across groups equal precision across groups
individual fairness	FTU/Blindness	-	no explicit use of sensitive attributes
	Fairness Through Awareness	-*	similar people are given similar decisions
causality-based fairness	Counterfactual Fairness	-	an individual would have been given the same decision if she had had different values in sensitive attributes
	path-specific Counterfactual Fairness	-	same as above, but keeping fixed some specific attributes

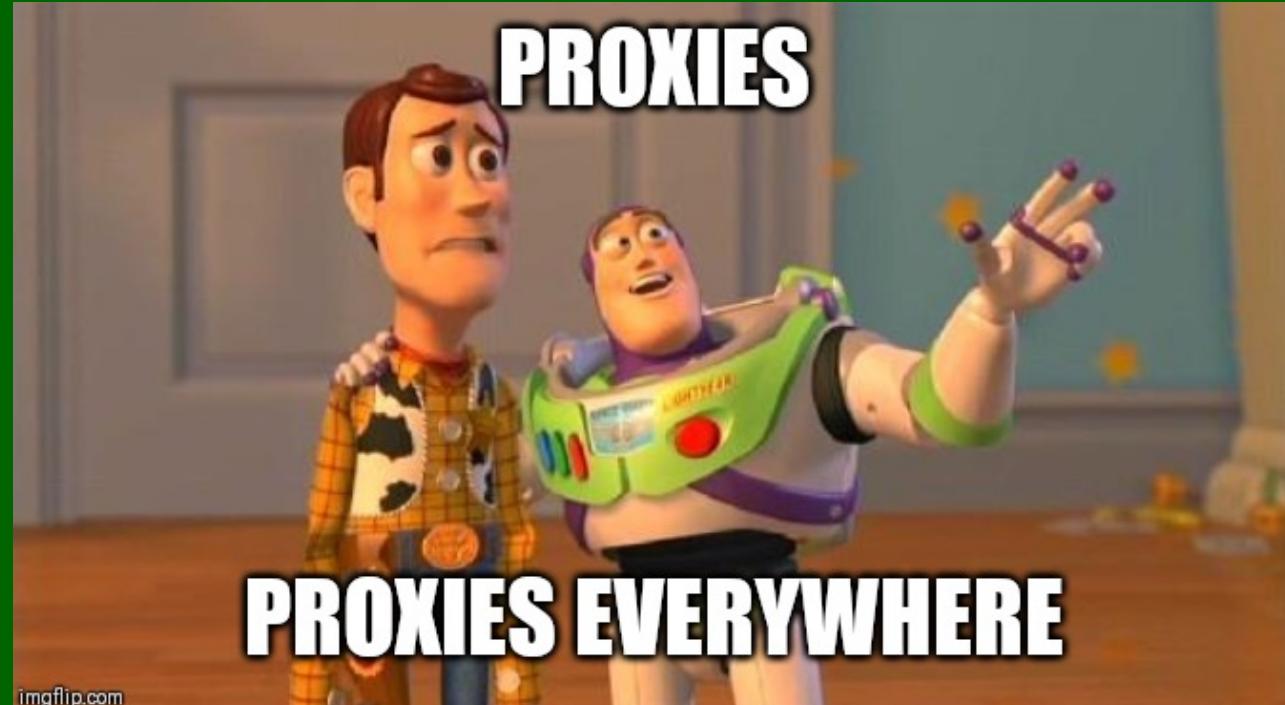
* there are exceptions to these cases where Y is actually employed, e.g. CDP conditioning on Y becomes Equality of Odds, and there are notions of individual fairness that use a similarity metric defined on the target space ([Berk et al., 2017](#)).

A3. Caso test médico

- Supongamos test genérico (con o sin técnicas estadísticas) de diagnóstico de una enfermedad. ¿Qué errores debemos minimizar?
- Respecto a la clase: ¿qué metricas nos interesan?
- ¿Y si el test requiere otra prueba invasiva y/o costosa?
- ¿Y si vamos a priorizar por riesgo de muerte a corto plazo?

A4. ¿Cómo podemos mitigar?

- Ya tenemos una medida del sesgo estadístico
- ¿Cómo podríamos mitigar?
- Pero antes: **¿tiene sentido una intervención estadística/algorítmica?**



Técnicas de mitigación de sesgos

UNDERSTANDING BIAS

Socio-technical causes of bias

- Data generation
- Data collection
- Institutional bias

Bias manifestation in data

- Sensitive features & causal inferences
 - Data representativeness
 - Data modalities

Fairness definition

- Similarity-based
- Causal reasoning
- Predicted outcome
- Predicted & actual outcome
- Predicted probabilities & actual outcome

MITIGATING BIAS

Pre-processing

- Instance class modification
 - Instance selection
 - Instance weighting

In-processing

- Classification model adaptation
- Regularization / Loss function s.t. constraints
 - Latent fair classes

Post-processing

- Confidence/probability score corrections
- Promoting/demoting boundary decisions
- Wrapping a fair classifier on top of a black-box baselearner

Fuente Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., ... Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), e1356. <https://doi.org/10.1002/widm.1356>

Caso detección paciente hepático: ILPD

Table 2 Experiment 3.1.1—unbalanced training data without feature selection, sex performance disparities

Mean difference averaged over n=100	Random forest classifier		Logistic regression classifier		Support vector machine		Gaussian Naïve Bayes	
	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p Value
Accuracy	2.96	0.00	-2.85	0.01	-2.98	0.02	-2.72	0.02
FScore	15.63	0.00	15.86	0.00	4.14	0.00	16.19	0.00
ROC_AUC*	6.80	0.00	2.93	0.00	-2.41	0.08	5.53	0.00
Precision	5.25	0.00	-4.87	0.00	3.41	0.00	-3.13	0.05
Recall	21.02	0.00	24.07	0.00	2.58	0.04	19.31	0.00
False negative rate	-21.02	0.00	-24.07	0.00	-2.58	0.08	-19.31	0.00
True negative rate	-7.42	0.00	-18.20	0.00	-7.40	0.00	-8.24	0.00
False positive rate	7.42	0.00	18.20	0.00	7.40	0.00	8.24	0.00
True positive rate	21.02	0.00	24.07	0.00	2.58	0.04	19.31	0.00

*ROC AUC score is a measure of the separation between classes in a binary classifier, derived from the area under the ROC curve.

"Across all classifiers females suffer from a higher false negative rate (FNR), while males suffer from a higher false positive rate"

Caso detección paciente hepático: ILPD

Table 5 Experiment 3.1.4—balanced training data with feature selection, sex performance disparities

	Random forest classifier		Logistic regression classifier		Support vector machine		Gaussian Naïve Bayes	
	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value	Sex performance disparities (%)	t-test p value
Accuracy	-5.62	0.00	-6.80	0.00	-6.19	0.00	-4.64	0.00
FScore	7.86	0.00	14.39	0.00	16.46	0.00	21.63	0.00
ROC_AUC	-0.05%	0.46	3.57%	0.00	5.95%	0.00	8.17%	0.00
Precision	4.60%	0.00	9.28%	0.00	12.82%	0.00	9.35%	0.00
Recall	9.70%	0.00	15.51%	0.00	15.38%	0.00	22.78%	0.00
False negative rate	-9.70	0.00	-15.51	0.00	-15.38	0.00	-22.78	0.00
True negative rate	-9.79	0.00	-8.37	0.00	-3.47	0.00	-6.44	0.00
False positive rate	9.79	0.00	8.37	0.00	3.47	0.00	6.44	0.00
True positive rate	9.70	0.00	15.51	0.00	15.38	0.00	22.78	0.00

"mixed results: the accuracy disparity benefits females across all classifiers, whereas the ROC_AUC disparity demonstrates a benefit for males in three out of four classifiers ... for all classifiers the FNR is consistently higher for females"

Herramientas ML para mitigación y explicabilidad



<https://fairlearn.org/>

Otras:

<https://ai-fairness-360.org/>

<https://pair-code.github.io/what-if-tool/>

Cuaderno Jupyter con FairLearn e ILPD

- Reproducción de los experimentos de Straw I, Wu H. BMJ Health Care Inform 2022
- Base de datos Indian Liver Patient Dataset (ILPD)

<https://github.com/javism/seminariofate2022/blob/master/IndianLiverPatientDataset-seminar.ipynb>

Auditando a un algoritmo interdisciplinamente

JUDGING THE ALGORITHM

\\

A case study on the risk assessment tool for gender-based violence implemented in the Basque country

Nov 17, 2022

Ana Valdivia (*Oxford Internet Institute*)
Cari Hyde-Vaamonde (*King's College London*)
Julián García Marcos (*Magistrate in Spanish Court*)

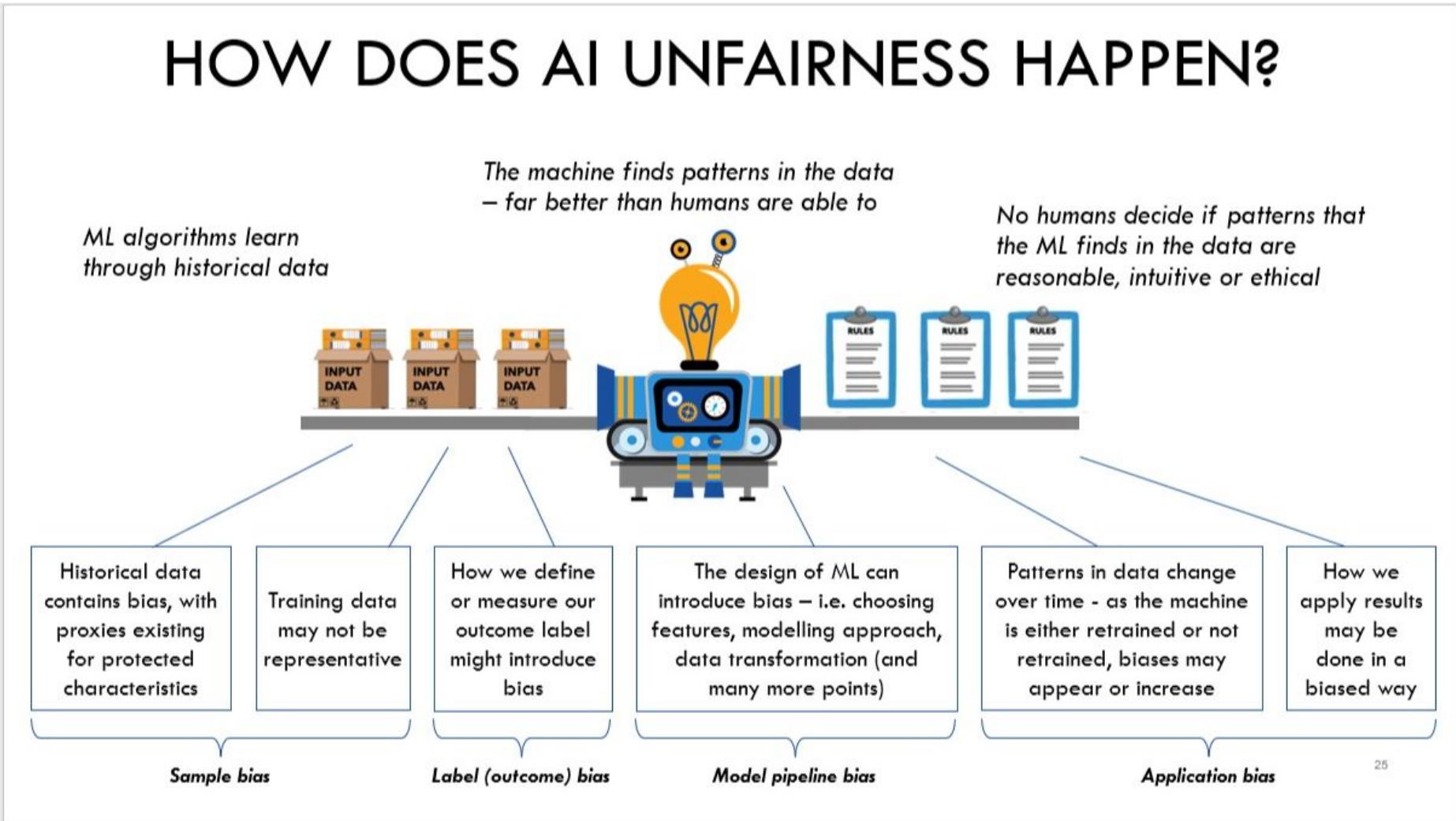


Enlace a la segunda parte Judging the algorithm (PDF)

A. Valdivia, C. Hyde-Vaamonde, J. García-Marcos. Judging the algorithm: A case study on the risk assessment tool for gender-based violence implemented in the Basque country. <https://arxiv.org/abs/2203.03723>

Resumen y Conclusiones

Recap: Fuentes de sesgo



Resumen

- El paso de prototipos de investigación a aplicaciones reales de la inteligencia artificial ha motivado la aparición de muchas áreas
- No solo FATE: IA robusta, privacidad en IA (aprendizaje federado, cifrado homeomórfico...), interacción persona-máquina (HCI)...
- Áreas implicadas según contexto: ética, derecho, política...
- Regulaciones (IA Act, GDPR, Ley Rider, AESIA...) y estándares (IEEE,ISO)
- Oportunidades de aprendizaje y comprender mejor los problemas y los conceptos de estadística.
- **¡¡Sistemas sociotécnicos!!**

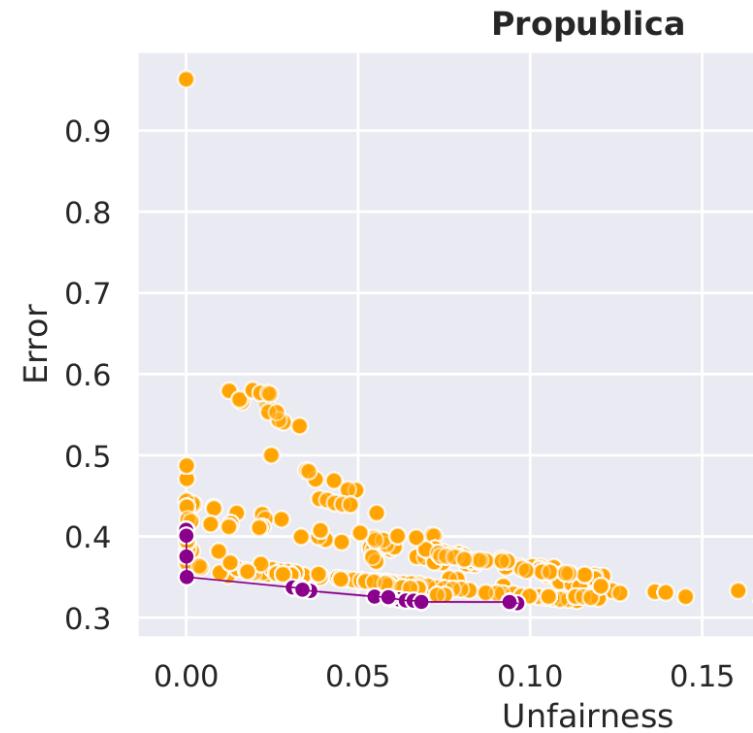
Trabajos relacionados de AYRNA

Explorar límites de precisión vs ecuanimidad

Valdivia, A., Sánchez-Monedero, J., & Casillas, J. (2021). How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness. *Int J Intel Sys*, 36(4), 1619–1643. <https://doi.org/10.1002/int.22354>

Índice alternativo al MELD/MELD-na

El grupo AYRNA en colaboración con el IMIBIC y otros centros trabaja en alternativas al MELD que no discriminan por género como estimador de riesgo de mortalidad en trasplantes hepáticos.



Trabajos relacionados de AYRNA

Desarrollo Ley Rider

Guía práctica y herramienta sobre la obligación empresarial de información sobre el uso de algoritmos en el ámbito laboral. *Ministerio de Trabajo y Economía Social. Gobierno de España.* 2022.

<https://prensa.mites.gob.es/WebPrensa/noticias/laboral/detalle/4125>

Proyecto AlgoRace

Proyecto AlgoRace. Investigación sobre discriminación racial e inteligencia artificial. 2021-2023. <https://algorace.org/>

Referencias (I)

- O'Neil, C (2018). Armas de destrucción matemática. Capitán Swing.
<https://capitanswing.com/libros/armas-de-destruccion-matematica/>
- Catherine D'Ignazio and Lauren F. Klein (2020). Data Feminism. MIT Press.
<https://mitpress.mit.edu/9780262044004>
- Solon Barocas and Moritz Hardt and Arvind Narayanan (2019). *Fairness and Machine Learning: Limitations and Opportunities*. <http://www.fairmlbook.org>
- Moritz Hardt (2020). *Fairness and Machine Learning* ([Part 1](#), [Part 2](#)) (MLSS 2020)
- Zhao, J. et. al (2017). Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. <https://www.aclweb.org/anthology/D17-1319>
- Buolamwini (2019). Response: Racial and Gender bias in Amazon Rekognition — Commercial AI System for Analyzing Faces.

Referencias (II)

- Verna, E. C., & Lai, J. C. (2020). Time for Action to Address the Persistent Sex-Based Disparity in Liver Transplant Access. *JAMA Surgery*, 155(7), 545–547. <https://doi.org/10.1001/jamasurg.2020.1126>
- Straw, I., & Wu, H. (2022). Investigating for bias in healthcare algorithms: A sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health & Care Informatics*, 29(1), e100457. <https://doi.org/10.1136/bmjhci-2021-100457>
- Castelnovo, A., Crupi, R., Greco, G. et al. A clarification of the nuances in the fairness metrics landscape. *Sci Rep* 12, 4209 (2022). <https://doi.org/10.1038/s41598-022-07939-1>
- Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., ... Staab, S. (2020). Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), e1356. <https://doi.org/10.1002/widm.1356>
- A. Valdivia, C. Hyde-Vaamonde, J. García-Marcos. Judging the algorithm: A case study on the risk assessment tool for gender-based violence implemented in the Basque country. <https://arxiv.org/abs/2203.03723>