Gender imbalance in medical imaging datasets for Artificial Intelligence

Prof. Christian Salvatore
University School for Advanced Studies IUSS Pavia

CEO and Co-Founder at
DeepTrace Technologies

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

Gender and sex differences in ML

- Machine learning (ML) (predictions by learning from data) is revealing a powerful emerging technology with well-documented results in improving screening, diagnosis and therapy and in defining biomarker signatures in several precision medicine applications,

- also where sex and gender differences has been reported

e.g. diabetes, cardiovascular, neurological, oncological diseases, and in immunology

Gender imbalance in medical imaging datasets for Artificial Intelligence

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

Adolf Hitler

Jean-Michel Basquiat

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

ML bias

- ML bias is a systematic error that occur in the ML model itself due to incorrect assumption in the ML process.
- ML depends on the quality, objectivity and size of training data. Faulty, poor or incomplete data will result in inaccurate predictions, reflecting the *garbage in – garbage out* admonishment used in computer science to convey the concept that the quality of the output is determined by the quality of the input.

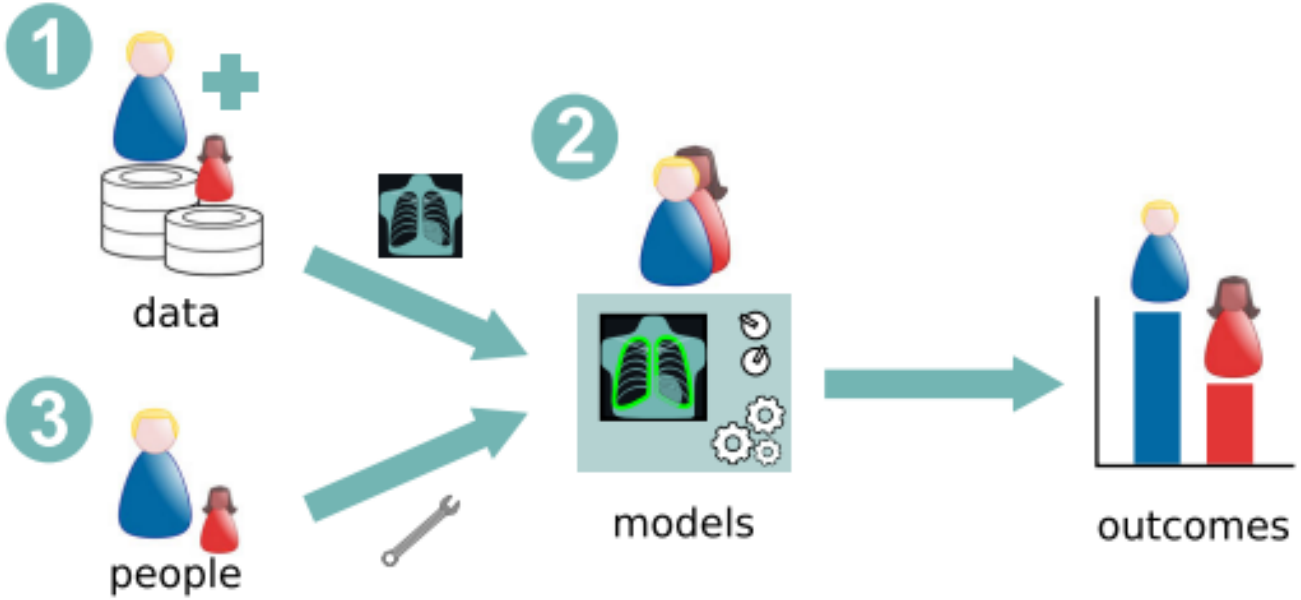
Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

Sources of ML bias

1. Data sample being fed to the ML model during training

2. Design choices for the ML model

3. People who develop the ML model



Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

Sample bias

- A problem with the data used to train the ML model.

- Data used is either not large enough or representative enough to train the system.

ML: the case of image recognition systems (convolutional neural networks, CNNs)

Image recognition systems that use biased machine learning data sets will inadvertently magnify that bias. Researchers are examining ways to reduce the effects.

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	PASTA
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	—
TOOL	KNIFE
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	GRILL
TOOL	TONGS
PLACE	OUTSIDE

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	VEGETABLES
HEAT	STOVE
TOOL	TONGS
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	MAN
FOOD	—
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

In this example of gender bias, adapted from a report published by researchers from the University of Virginia and the University of Washington, a visual semantic role labeling system has learned to identify a person cooking as female, even when the image is male.

Emergency Radiology (2022) 29:365–370
https://doi.org/10.1007/s10140-022-02019-3

ORIGINAL ARTICLE

Check for updates

Deep learning prediction of sex on chest radiographs: a potential contributor to biased algorithms

David Li[1,2] · Cheng Ting Lin[3] · Jeremias Sulam[4] · Paul H. Yi[2]

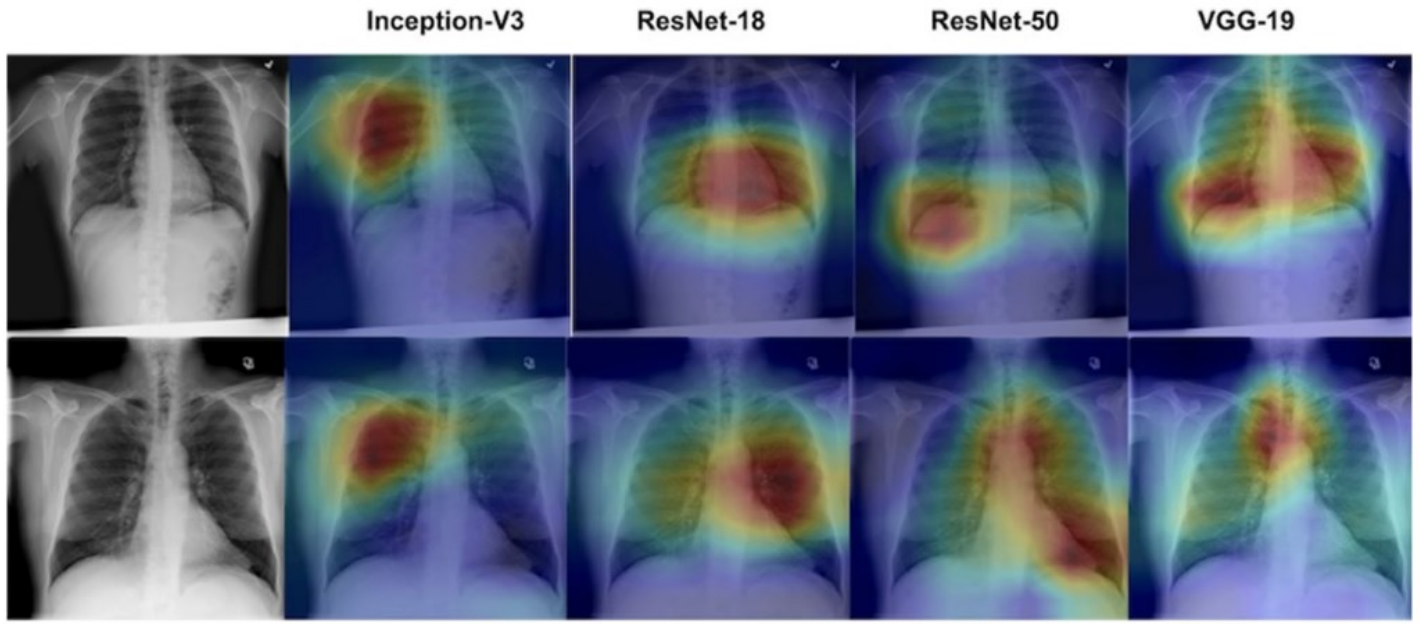


Fig. 1 Heatmaps for predicting sex demonstrate similar activations across different CNN architectures

A convolutional neural network (CNN) trained on two large chest radiography datasets accurately predicted sex with similar localizations (heatmap) across different architectures and datasets.
These DCNNs can be beneficial to emergency radiologists for forensic evaluations.
On the other hand, these findings support the notion that CNNs can potentially confound the accurate prediction of disease and contribute to biased models.

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

Sex/gender bias in CNNs

- In general, it is well known that image recognition systems (convolutional neural networks, CNNs) tend to learn representations useful to solve the task they are being trained for.

- When we go from male to female images (or vice versa), structural changes in the images appear, leading to a change in data distribution which may be associated to a decrease in predictive performance.

- Algorithmic solutions to such “domain adaptation” should be engineered, especially in cases when it is difficult to obtain gender-balanced datasets

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

ML desirable and undesirable bias

- if desirable bias may be accepted by including in ML modelling sex or gender features to highlight sex/gender-based differences in predictions.
- undesirable biases should be avoided as those derived from training ML models on datasets with under-representation of sex/gender minority groups.

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

ML desirable and undesirable bias

- It is important to warrant that ML systems do not generate or widen sex- and gender-based disparities in access to healthcare (group equity)
- It is however important that ML systems model sex and gender differences when differences exists, e.g. in risk factors of disease developing or progressing (individual equity)

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

ML desirable and undesirable bias

- Otherwise, if there are intrinsic differences in the population, such as sex or gender differences in disease prevalence, ML models that well suit to the majority group may not be generalized for minority ones.

- Sex and gender differences should be taken into consideration in ML algorithms in those clinical applications where there are important differences in the epidemiology and clinical presentation of conditions.

12592–12594 | PNAS | June 9, 2020 | vol. 117 | no. 23

Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis

Agostina J. Larrazabal[a,1], Nicolás Nieto[a,b,1], Victoria Peterson[b,c], Diego H. Milone[a], and Enzo Ferrante[a,2]

[a]Research Institute for Signals, Systems and Computational Intelligence sinc(i), Universidad Nacional del Litoral–Consejo Nacional de Investigaciones Científicas y Técnicas CONICET, Santa Fe CP3000, Argentina; [b]Instituto de Matemática Aplicada del Litoral, Universidad Nacional del Litoral–Consejo Nacional de Investigaciones Científicas y Técnicas, Santa Fe CP3000, Argentina; and [c]Facultad de Ingeniería, Universidad Nacional de Entre Ríios, Oro Verde CP3100, Argentina

A decrease in performance for under-represented genders was found when a minimum balance is not fulfilled.

Computer-assisted diagnosis systems should include explicit gender balance and diversity recommendations.

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

Fairness

Fairness in ML implies the application of strategies for preventing or correcting bias in automated decision processes based on ML predictive models.

Decisions made by a model after a ML process may be considered unfair if they are based on data features considered sensitive.

Examples of these kinds of variable include sex and gender.

Table 1 | Databases commonly used in fairness in MIC studies

Image modality	Database	Access	Sex or gender[a]	Age	Skin tone or race/ethnicity[b]	SES
Chest X-ray	CheXpert[31]	Public	x	x	x	–
	NIH Chest X-Ray[32]	Public	x	x	–	–
	MIMIC Chest X-Ray[33]	Public	x	x	x	x
	Emory University Hospital Chest X-Ray[20]	Private	x	x	x	–
Mammography	Digital Mammographic Imaging Screening Trial (DMIST)[34]	Private	x	x	x	–
	Emory University Hospital Mammography[20]	Private	x	x	x	–
Dermoscopy	ISIC Challenge 2017/18/20[35,36]	Public	x	x	–	–
Dermatological clinical image	Fitzpatrick 17k[13]	Public	–	–	x	–
	SD-198[49]	Public	–	–	–	–
Fundus image	AREDS[37]	Public	x	x	x	–
	Kaggle EyePACS[50]	Public	–	–	–	–
Cardiac MRI	UK Biobank[38]	Public	x	x	x	x
Pulmonary angiography CT	Stanford University Medical Center[16]	Public	x	x	x	–

[a]According to the World Health Organization, *sex* refers to different biological and physiological characteristics of males and females, while *gender* refers to the socially constructed characteristics of women and men such as norms, roles and relationships of and between groups of women and men. Databases tend to report one or the other.

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

Criteria of fairness

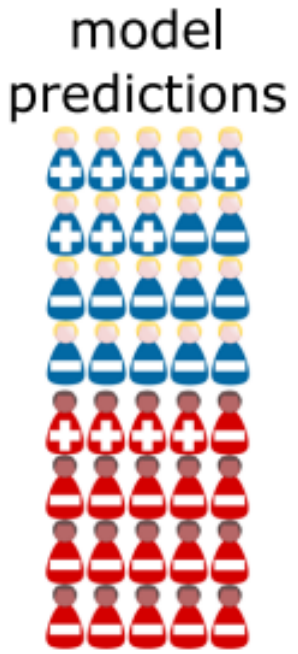
A recent explosion of research has investigated different criteria of fairness and how to ensure that they are satisfied by predictive algorithms.

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

ML statistical fairness metrics

Statistical Parity (predictive parity):

The (expected) percentage of individuals predicted to be positive by ML model is the same for each relevant group.



Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

ML statistical fairness metrics

The statistical fairness metrics are provably inconsistent except under very specific conditions that are in practice unattainable.

The goal of complete sex/gender neutrality is unachievable.

At best, we can make optimal trade-offs between different kinds of unfairness.

Kleinberg et al., 2016; Chouldechova, 2017; Miconi, 2017

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

(ML) statistical fairness metrics

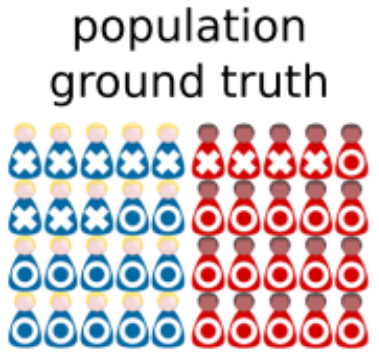
Importantly, the criteria of fairness and the impossibility theorems apply to all predictions, whether generated by machine learning, questionnaires, human judgment, or any other mechanism.

Kleinberg et al., 2016; Chouldechova, 2017; Miconi, 2017

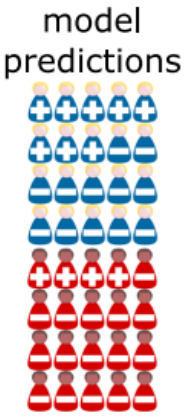
Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

The paradox

Suppose that, for women and men patients, the prevalence of the condition differs between their respective groups.



Then suppose that the model satisfies predictive parity for women and men patients.

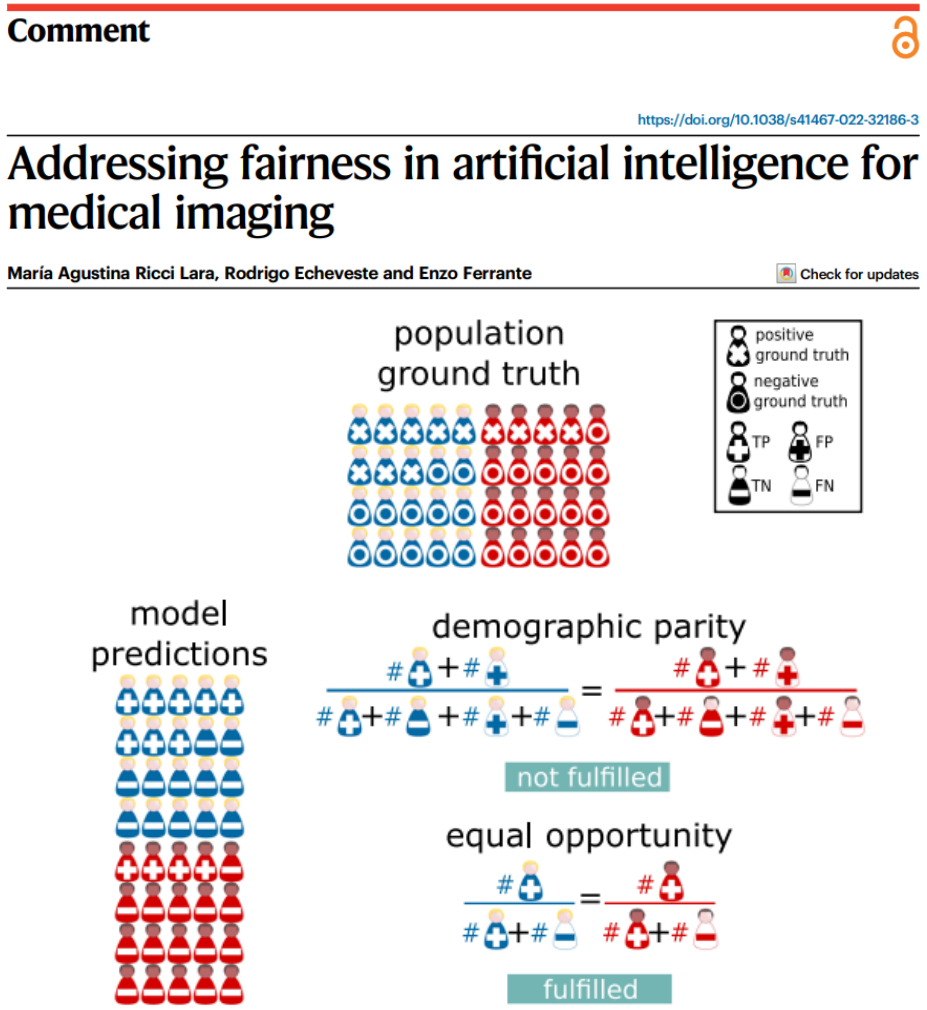


The problem is that under these conditions, either the algorithm's predictions are perfect, or the false positive and false negative rates differ for women and men patients. This is a straightforward consequence of the differing base rates across both populations.

The paradox

The problem is that under these conditions, either the algorithm's predictions are perfect, or the false positive and false negative rates differ for women and men patients.

This is a straightforward consequence of the differing base rates across both populations.



Comment

https://doi.org/10.1038/s41467-022-32186-3

Addressing fairness in artificial intelligence for medical imaging

María Agustina Ricci Lara, Rodrigo Echeveste and Enzo Ferrante

Fig. 1 | Group-fairness metrics. Here we include a toy-example in the context of disease classification, where two sub-populations characterized by different protected attributes (in red and blue) present different disease prevalence (40% and 20% for blue and red subjects respectively, top row, **x** marks positive cases). A model optimized for discriminative performance was assessed on a test set

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

An example: breast cancer

Although breast cancer is much more common in women, men can develop breast cancer.

In the United States, less than 1% of all breast cancers occur in men.

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

ML trade-offs

When selecting an appropriate statistical fairness metric, ML developers will need to balance different trade-offs.

Is it more important to focus on false positive or false negative diagnoses in clinical environments?

Biddle, 2020

ML fairness and final fair decision

Algorithmic Fairness and Final Fair Decisions

 Philosophy & Public Affairs

BRIAN HEDDEN*

On statistical criteria of algorithmic fairness

Unfair algorithmic decision-making relative to certain statistical criteria does not imply unfair final decision-making.

Biases exhibited by ML algorithms can be identified and corrected for in the human part of the decision-making process.

ML fairness and final fair decision

Ethics and Information Technology (2022) 24:39
https://doi.org/10.1007/s10676-022-09658-7

ORIGINAL PAPER

Check for updates

Enabling Fairness in Healthcare Through Machine Learning

Thomas Grote[1] · Geoff Keeling[2]

This study examines how the interplay of a clinician and an algorithm, that overfit (performs worse) for disadvantaged patient groups, might result into fair final decisions for all the patients.

Useful for ML as decision support systems, that are the most frequent applications in medicine.

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

Affirmative ML

- Algorithms trained on diverse datasets that perform better for some groups can be permitted under specific conditions.

- The balanced properties of algorithmic decisions is not the most important factor in the decision.

- Whilst such algorithmic decisions may be unfair, the fairness of algorithmic decisions is not the appropriate locus of moral evaluation.

- What matters is the fairness of final decisions, such as diagnoses, resulting from collaboration between clinicians and algorithms.

- Affirmative algorithms can permissibly be deployed provided the resultant final decisions are fair.

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

Collaborative ML

Alternative forms of collaboration between ML algorithms and clinicians can promote fairness in healthcare, even if the ML algorithm is biased.

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

1. The peer model

ML algorithms and clinicians offer competing predictions about (for example) the correct diagnosis or the best treatment recommendation.

The salient feature of this approach is that clinicians and algorithms address the same clinical task, and their solutions to that task is balanced.

ML model is a peer!

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

2. The triage model

The idea is that an algorithm's prediction may be causally upstream of the clinician's judgement such that the clinician's judgement is enhanced in virtue of the clinician's knowledge of the prediction.

ML as an instructor!

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

The division of the labour

What motivates the division of labour is the observation that there is an opportunity-cost to clinicians performing any clinical task

The benefit of this model of clinician-algorithm collaboration is that, in specific cases, the algorithm alleviates the burden on clinicians given that expert clinicians in the relevant domain are scarce.

In specific cases, in which the confidence of the clinician falls below a pre-defined threshold and in which the algorithm has high confidence, the clinician defers to the algorithm.

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

3. Hybrid ML

A way to accommodate bias preservation might be by assigning a hybrid role to the algorithm, in which it acts both as a peer and an instructor.

In specific cases, in which the confidence of the clinician falls below a pre-defined threshold and in which the algorithm has high confidence, the clinician defers to the algorithm.

In specific cases, in which the confidence of the clinician falls over a pre-defined threshold and in which the algorithm has less confidence, the clinician does not defer to the algorithm.

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

3. Hybrid ML

It is however recommended:

- supplementing ML algorithm with an explainable result to clinician (e.g. heatmap, morphological and texture predictors)

- additional safeguards need to be implemented. In particular, clinicians might be required to diagnose a fraction of male patients themselves (even if they are less confident than the algorithm), while receiving further supervision.

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

Conclusions

- Not enough large or representative dataset to train a ML system can be a source of ML bias.
- Undesirable biases should be avoided as those derived from training ML models on datasets with under-representation of sex/gender minority groups, since they could affect the confidence of ML in these groups.
- However, the goal of complete sex/gender neutrality could be unachievable. At best, we could make optimal trade-offs between different kinds of unfairness.

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

Conclusions

- Overemphasis on the evaluative properties of ML algorithmic decisions is likely to hinder efforts to rectify injustices.

- The processes by which ML algorithms and clinicians jointly contribute to final decisions is a possible solution and could be defined to determine exactly how algorithms can be designed and deployed in a way that the final collaborative decision promotes equity in health outcomes.

Gender imbalance in medical imaging datasets for Artificial Intelligence

Prof. Christian Salvatore
IUSS Pavia
CEO and Co-Founder at DeepTrace Technologies

7. SÜDTIROLER SYMPOSIUM
live vor Ort und digital

7° SIMPOSIO ALTOATESINO
in presenza e online

christian.salvatore@iusspavia.it