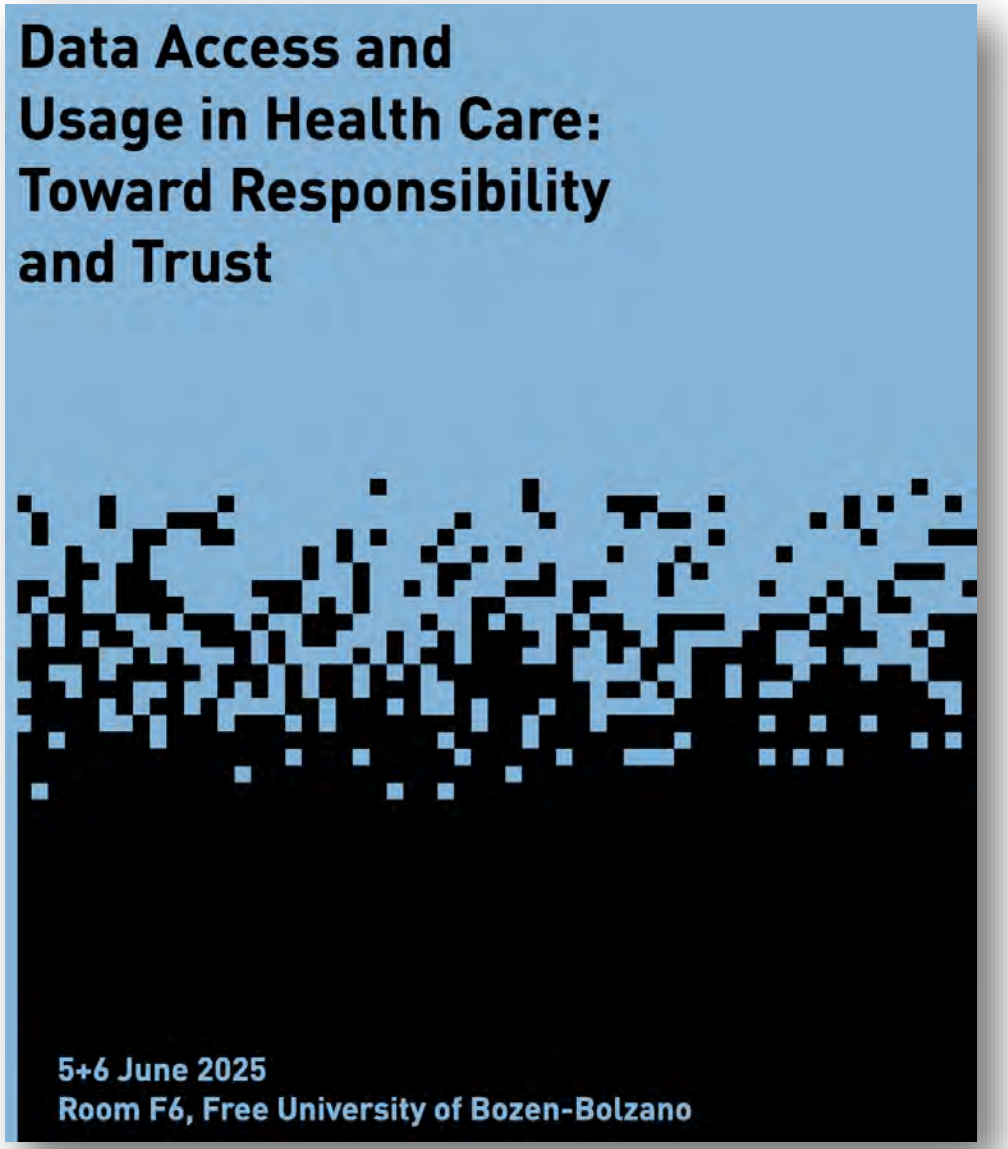


On the AI Ethics Principle of Fairness in Healthcare Machine Learning

Simona Tiribelli

University of Macerata, Italy

Institute for Technology & Global Health, USA

What does it mean to develop AI systems in healthcare taking into account the **ethical principle of fairness**? What does it entail to operationalize or embed «fairness» through (by design) AI systems used in healthcare?

Context

- Responsible data access and usage in healthcare => **fairness as a central concept** (and principle) before AI ethics in data ethics, research ethics, bioethics, medical ethics, etc.
- AI models increase the complexity of the issue per se (from data ethics to AI ethics) => from fair data collection and usage to fair (design [data – variables], development, and use of) AI models that are more and more complex and very often opaque ("black box" – Pasquale 2015)
  - **Benefits**: > prediction (probabilistic) -- crucial in healthcare (e.g., pathology occurrence and development in clinical domain, outbreak detection in public health, etc.); > accurate
  - Capacity to discover hidden pattern by processing huge amounts of multimodal health data is of paramount importance
  - **!!! Opaque** – even more with DL and LLMs – due to AI complexity (not just trade secret) = inscrutable
- …especially when trained on healthcare data and developed and deployed for more and more critical tasks in healthcare and medicine (i.e., access to and distribution of health and care resources/services; detection, prediction, and management of health and clinical pathologies; clinical logistics; etc.)
- Understanding how fairness is understood and operationalized (i.e., what concept of fairness is guiding mainly AI research and practice, i.e., design and use of AI in healthcare) is clearly of paramount importance – especially bearing in mind the idea of using technology for improving societies and healthcare ecosystems.

Goals

1. Shed light on how fairness is mainly understood in healthcare AI (and ethics)
2. Assess whether this concept is adequate in order to design and deploy truly fair AI systems
3. Propose how fairness – as an AI ethics principle – can be enriched drawing insights from moral philosophy and ethical theory
4. Highlight practical steps to operationalize the AI ethics principle of fairness in healthcare AI

Agenda

I. Fairness in (healthcare) AI

II. The limits of the current conceptions of fairness in (healthcare) AI / AI and ethics debate

III. On the ethical value of fairness – what fairness truly demands

IV. The path toward “substantial fairness” in (healthcare) AI

Agenda

I. Fairness in (healthcare) AI

II. The limits of the current conceptions of fairness in (healthcare) AI / AI and ethics debate

III. On the ethical value of fairness – what fairness truly demands

IV. The path toward “substantial fairness” in (healthcare) AI

I. Fairness in Healthcare AI

- Fairness as a «core topic» in the debate on AI and ethics in healthcare
- Cases of "unfair AI" abound in many sectors [criminal justice, education, **healthcare,** e.g., assign priority in access to special resource programs; decide on hospitalization; cancer detection and pathology development prediction, calculation of insurance rate, etc.]
- Fairness as a «necessity» in academia and industry [*reactive approach*]
- **Lowest Common Denominator:** biased AI models

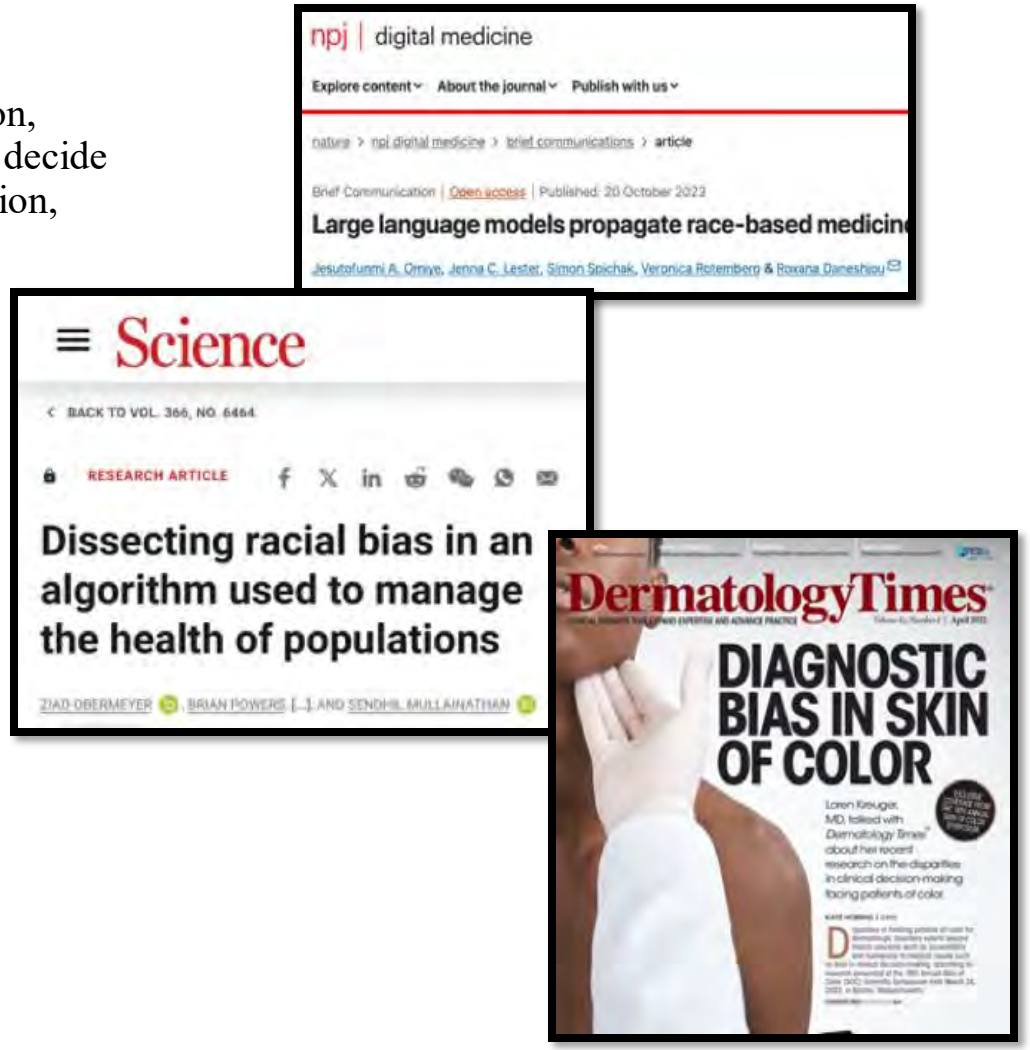
  → biased outputs

  → against certain groups

"already vulnerable or historically marginalized"

→

- Unfair decisions on health matters
- Unfair (health – social) treatment
- Unfair access to (health – social) services and care



I.1 Fairness in Healthcare AI

- ➢ «Core topic» in the debate on AI and ethics («responsible and trustworthy AI»)
- Core «AI ethics principle» in the more than 200 AI ethics frameworks developed to date for responsible and ethical AI at the global level (see Jobin et al. 2019, Correa et al. 2023)
  - by business / private sectors (self-assessment tool for responsible AI – see Microsoft, IBM, etc.)
  - by governments and international institutions (see EU commission, *Ethical Guidelines for Trustworthy AI*; WHO *Ethics and Policy for AI in Health* 2021-2023; etc.)

- ✓ Fairness as a «requirement» in core in regulatory and strictly normative regulations [see **EU AI Act** 2024)
- ✓ Convergence on how fairness in AI and ethics (trustworthy AI debate) is mainly understood and hence addressed

I.2 Fairness in Healthcare AI

- Fairness in AI and ethics debate or "FAIR AI" as "***non-biased AI"*** or "***bias-free" AI***

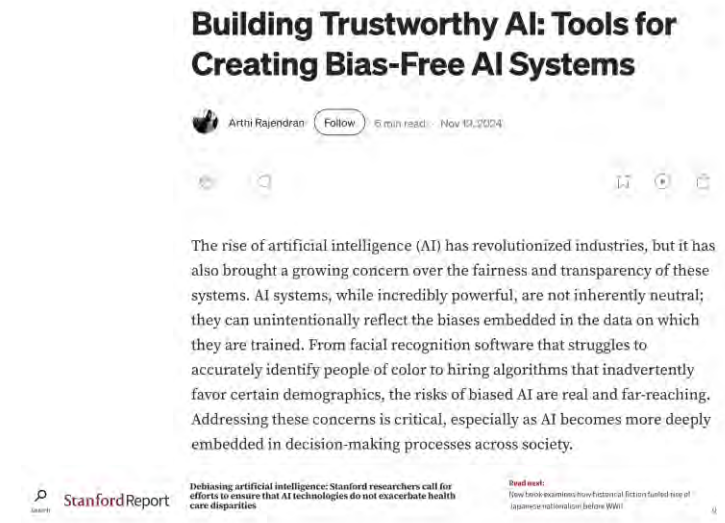
- *Fairness as a response to algorithmic discrimination, as algorithmic discrimination is linked to bias in AI, major focus on bias detection and removal to create fair AI*

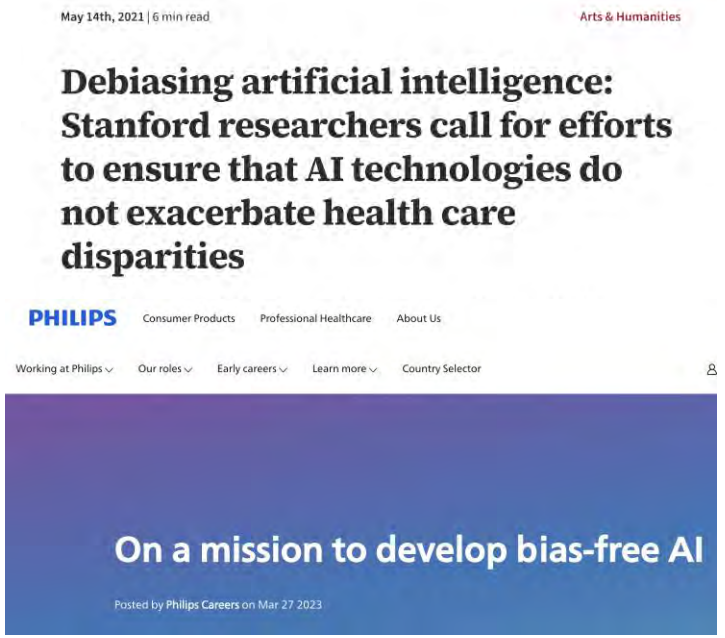
- Eliminating biases => debiasing AI => fair AI

- Biases related to protected attributes (i.e., gender, ethnicity, etc.) especially in datasets and models

- Solutions: «**bias as technical bugs» = bias detection and removal techniques (e.g., removal of sensitive attributes – see def. of fairness as "anti-classification")**

  - Not always possible, even counterproductive (e.g., in healthcare: biological variations between genders can influence the effectiveness of certain medications)
  - Not successful or relevant to design fair AI (see Obermayer et al. 2019)





Agenda

I. Fairness in (healthcare) AI

II. The limits of the current conceptions of fairness in (healthcare) AI / AI and ethics debate

III. On the ethical value of fairness – what fairness truly demands

IV. The path toward “substantial fairness” in (healthcare) AI

II. The limits of the current conceptions of fairness in (healthcare) AI

1ST LIMIT

- Bias as a 'technical bug' to be eradicated = **oversimplified approach**
- Bias as "socio-technical patterns of longstanding and present inequalities"
- Systems re-learn such bias (biased epistemic sources and from biased users = e.g., bias in benchmark medical methods and biased medical beliefs)
- "ideal systems" for "non-ideal societies"
- Perpetuation of systemic inequalities/*status quo*

Biases are not merely mathematical bugs and technical errors to fix. Rather, gender and ethnic biases – as morally unfair and harmful correlations – reflect the hidden patterns of systemic oppression and structural inequalities that historically unfairly affect and shape healthcare access, quality of healthcare, and benefits healthcare substantial enjoyment fruition. To put it differently, such biases are social determinants of health and healthcare (Paradies, Bastos, and Priest 2015; Ramsoondar, Anawati, and Cameron 2023). Their acritical or passive removal in healthcare AI development is – and has been shown to be – neither useful nor effective – missed societal opportunity (Tiribelli 2025, in «Ethics and Medical AI», Springer, 2025, in press)

II.1 The limits of the current conceptions of fairness in (healthcare) AI

"Gender and racial biases are indeed encoded and perpetuated through medical knowledge, clinical methods (including benchmark scales), traditional tools (e.g., pulse oximeter medical devices), and established health practices, as well as via clinician evaluations and encounters, physician-patient communications, and individual beliefs and attitudes (Diao et al. 2021; Smedley, Stith, and Nelson 2003; Sudat et al. 2023) – as in the case of marginalized people who have internalized prejudices and oppression in their health and care perspectives and judgments (Veltman and Piper 2014). For example, half of white medical trainees still believe such myths as black people have thicker skin or less sensitive nerve endings than white people – biased, often unconscious misbeliefs that lead to unfair medical treatment and healthcare quality for the considered group (Hoffman et al. 2016)" (Tiribelli, *Toward Healthcare (Social) Justice: On the Value of Biases in Healthcare AI*, 2025)

II.2 The limits of the current conceptions of fairness in (healthcare) AI

2ND LIMIT

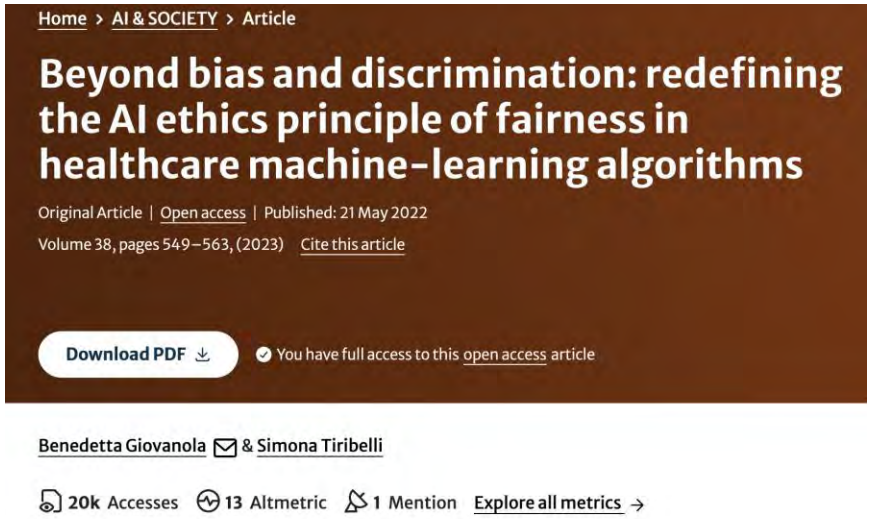
Formalization/CONCEPTUALIZATION problem:

- FAIR MODELS = "Parity models"/equalized odds: fairness is formalized as **same model performance for all groups (AD/DV)** = fairness is just measured as disparity in performance and outcomes between advantaged or disadvantaged groups (*mathematical understanding of fairness*)

- Algorithmic fairness entails adjusting the models in order to reduce gaps in performance and outcomes between diverse demographic groups, hence ignoring other social and contextual factors.

✓ **Problems:**

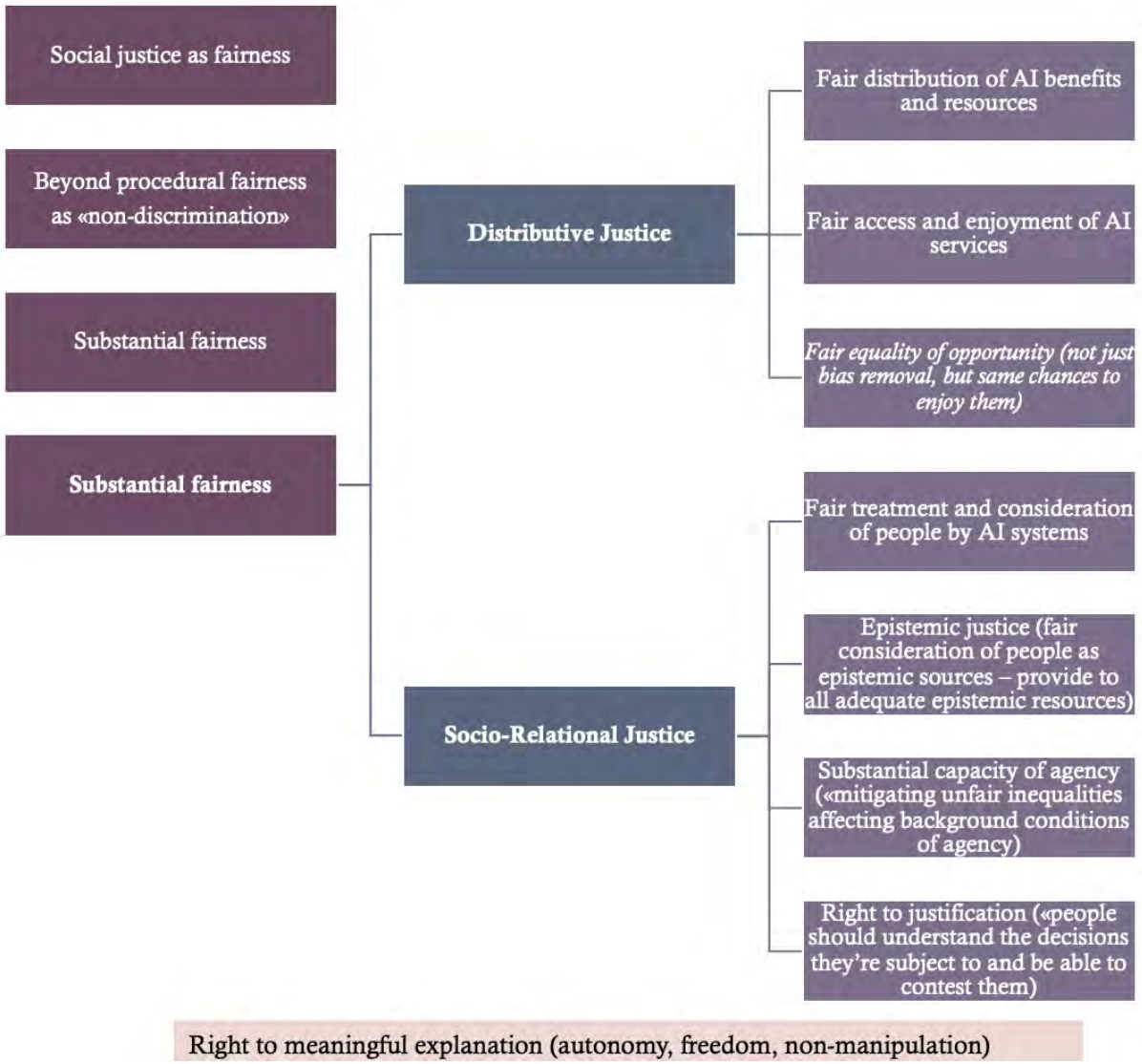
✓ Practical (effect): Performance degradation / «levelling down effects» (Mittelstadt, Wachter, and Russell 2023)
✓ Theorethical (cause): thin concept of fairness (more mathematical, than ethical): procedural fairness («formal equality of opportunity») "negative concept of fairness" = as "non-discrimination"



!!! Adjusting the model does not fix the source of the poblem;
!!! Does not work and does not ensure fairness in contemporary societies (imbued with inequalities)
!!! Fairness as an ethical value demands more that same solutions for all:
it entails the substantial promotion of opportunities for all, as real chances for every person to express their agency and therefore the development of adequate conditions for people to afford them (*substantial fairness*) with a specific regard to the worst off.

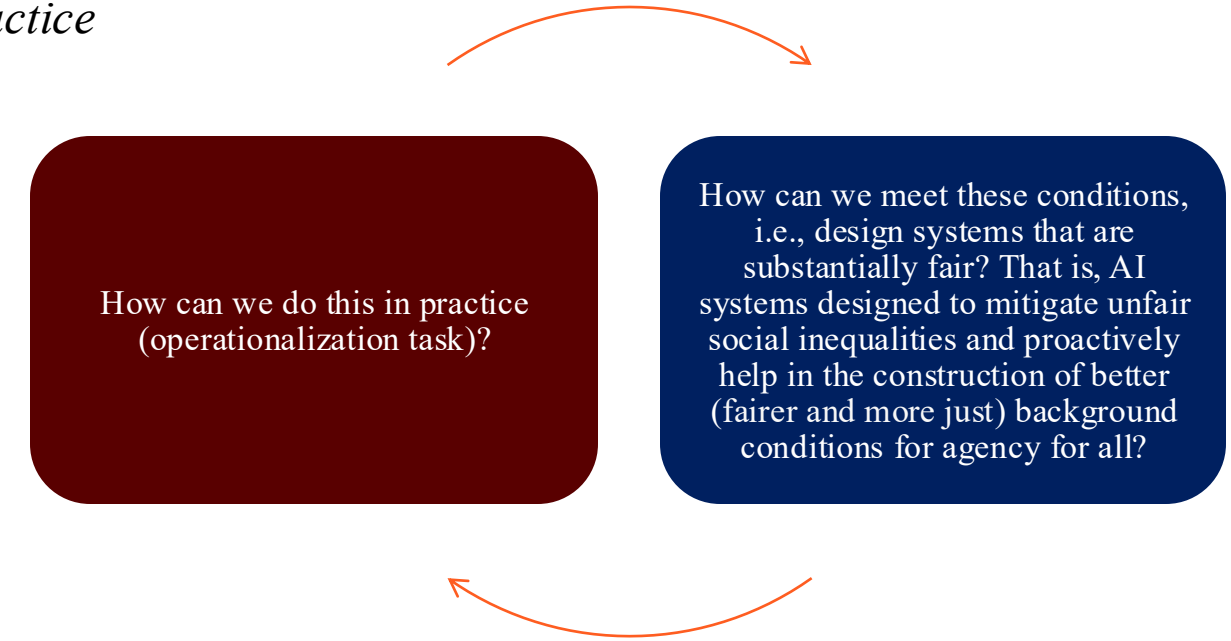
III On the ethical value of fairness

— *what fairness truly demands*



IV. The path toward “substantial fairness” in (healthcare) AI

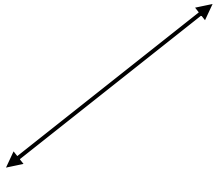
- What to do: from *theory* to *practice*



IV. The path toward “substantial fairness” in (healthcare) AI

- What to do: from *theory* to *practice*

- Non-idealized and stereotyped samples
- People can have unconscious biases
- People can internalize oppression and prejudices



AI substantial fairness methodological pipeline:

1. A proactive and critical view on bias («bigger picture» - zoom in)
2. «Decoding Bias»
3. «AA action» by design

+ inclusive and representative datasets of non-ideal, real users

+ heterogeneous and properly trained design team (specific training on substantial fairness)

+ real users involvement into the whole AI lifecycle (e. g. focus groups)

IV. From *theory* to *practice:*

1. BIAS: zoom in «bigger picture»

Beyond technical gaps to fix: bias as a "socially situated, multi-relational phenomenon" (zoom in the focus)

Investigate the social context and relations of AI development, deployment, and use

Beyond removing bias through sensitive attributes or post processing adjustments of the model, raise deeper questions on the relation between data – model – people in a context sensitive way?

Label bias

Variable bias

Cohort bias

Target bias

What is the background of who is gathering and labeling data used to train or fine-tune AI models? What physical, psychological, cultural factors influence that activity? Is that accurate? How can we measure the accuracy? Is this information transparently communicated?

Who are the people that can or may use these systems? Have been involved in the whole AI development and test phase and their inputs adequately heard? Who might be accidentally excluded from the design of this technology? Why? Who are the indirect users or subject affected by the system? Who might be the (unconscious and conscious) biases affecting them?

Missing data bias

Minority bias

Informativeness bias

Training-service skew

Epistemic bias

Automation bias

Bias of feedback loop

Dismissal bias

What are the asymmetries in knowledge and action that characterize the users of AI solutions? What lack of physical, epistemic, and economic resources might affect the fair access and use of AI solutions? Can we address them by design?

Epistemic bias

Privilege bias

Informed mistrust bias

Agency bias

IV. From theory to *practice*

2. Decoding bias

✓ **«Decoding Bias» -> «discrimination debunking»**

- ❑ Harnessing AI capacity to process multimodal data => discover patterns of inequalities
- ❑ Considering and inquiring them «patterns of [unfair] inequalities» / determinants of injustice
- ❑ Design AI systems considering them to ensure AI systems work promoting fairness
- ❑ Debunking discrimination in the human social dimension
- ❑ Use that information to re-design or fine-tune AI models

(For example, Cindy Zhang, Sarah Cen, and Devavrat Shah propose algorithm models that can identify discriminatory samples in the data that could cause disparities in systems' predictions, then re-used to retrain DL models for more fair predictions – see *Matrix estimation for individual fairness* 2023)

- ❑ AI addressing the social root of the problem

IV. From theory to practice – 3. AA by design (?)

AA by design

"A policy, an act, or an intervention (etc.) is an AA if and only if it prescribes and enacts positive steps to increase the representation of women and minorities in a relevant area from which they have been historically excluded and/or aims to reasonably address the disadvantages they suffer in some ways other than boosting their representation [(e.g., through quotas)]"
(Lippert-Rasmussen 2020)

Numerical AA

- Setting quotas for gender and ethnic minorities, etc. (also known as strong AA)

→ «Levelling up» (Watcher, Mittelstadt, Russel, 2023) increasing by design decision rates or recalls to the required level for those groups usually harmed by decision rates or recalls that are too low

Non-numerical AA

- E.g., the cases of outreach AA programs involving special efforts to advertise open positions to underrepresented groups or remedial training programs (also known as weak AA)

→ "Compensatory options / action by design" – harness bias to mitigate them!

IV. – an example

Consider a healthcare AI decision-making system designed to support clinicians or HCPs to determine priority in access to first aid facilities, designed using as a proxy patients' reported pain levels.

In this context, historical data show that women and individuals from ethnic minority groups often have their pain underestimated compared to white men (Hoffman et al. 2016). If the AI system is solely trained on historical data, it will easily inherit and perpetuate these biases, leading to a biased prioritization process, widening existing inequalities by negatively affecting historically deprived categories.

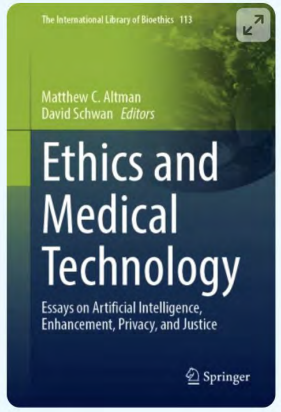
Such bias still depends on misbeliefs that are very common among clinical trainers and residents; therefore, even if the system would be trained on controlled (deprived of biases) data samples, by interacting with clinicians and patients will easily up re-learning them due to automation bias.

To prevent this risk, the AI model could be designed by intentionally leveraging sensitive attributes such as gender and ethnicity, utilizing them as decision nodes to identify and signal the possible risk of underestimation or inadequate treatment for patients from minorities and historically disadvantaged groups.

For instance, if the system assesses a Black woman reporting severe pain, it can signal to the clinician and patient that this is an 'at-risk demographic' for the specific task to evaluate: the system would flag this patient as belonging to a demographic group that faces a higher likelihood of pain underestimation, based on reported scientific evidence on historical biases in this regard over time. The systems can be trained using flags for second-order proxy for the possible need for compensatory tools.

The systems can then display compensatory tools, such as additional options whose outputs can be introduced in the assessment, weighted along with others, so as to calibrate the model toward a more accurate and fair decision for the specific patient. Such compensatory tools can be, for instance, semi-structured assessments based on diverse parameters and indicators. This is very relevant because, for example, the experience of pain *differentially* activates stress-related physiological responses across various ethnic groups and members of different ethnic groups show to use differing coping strategies in managing pain complaints, which therefore need to be considered (Campbell and Edward 2017).

Home › Book



Conclusive remarks

- Fairness as a complex ethical issue (multidimensional concept) in healthcare AI
- Beyond «negative concept», «positive» or «substantial» concept of fairness in healthcare AI
- Beyond removing «bias», **«decoding bias»** to **«debunk and dismantle discrimination»**
- Avoid «perpetuating status quo», using AI (through dynamic soft AA by design) to compensate unfair background conditions of «agency»
- Beyond FAIR as non-biased AI systems, FAIR AI systems as systems capable to promote better, more just and fairer, societies and healthcare ecosystems.

THANK YOU

Simona Tiribelli, PhD

UNIVERSITY OF MACERATA | simona.tiribelli@unimc.it