Responsible AI

Introduction to Machine Learning in Healthcare

Nuno André da Silva
Jorge Cerejo

HOSPITAL DA LUZ
LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Questions from last week?

HOSPITAL DA LUZ
LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

What is responsible AI ?

HOSPITAL DA LUZ
LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Why do we need responsible AI ?

HOSPITAL DA LUZ
LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

AI's Impact in Society

AI can significantly impact society due to its large range of applications

Some dimensions

- **AI & Democracy**
- **AI & Future of work**
- **AI & Rights equality**
- **AI & Medicine**

https://unesdoc.unesco.org/ark:/48223/pf0000389736

AI's Impact in Society

AI can significantly impact society due to its large range of applications

Some dimensions

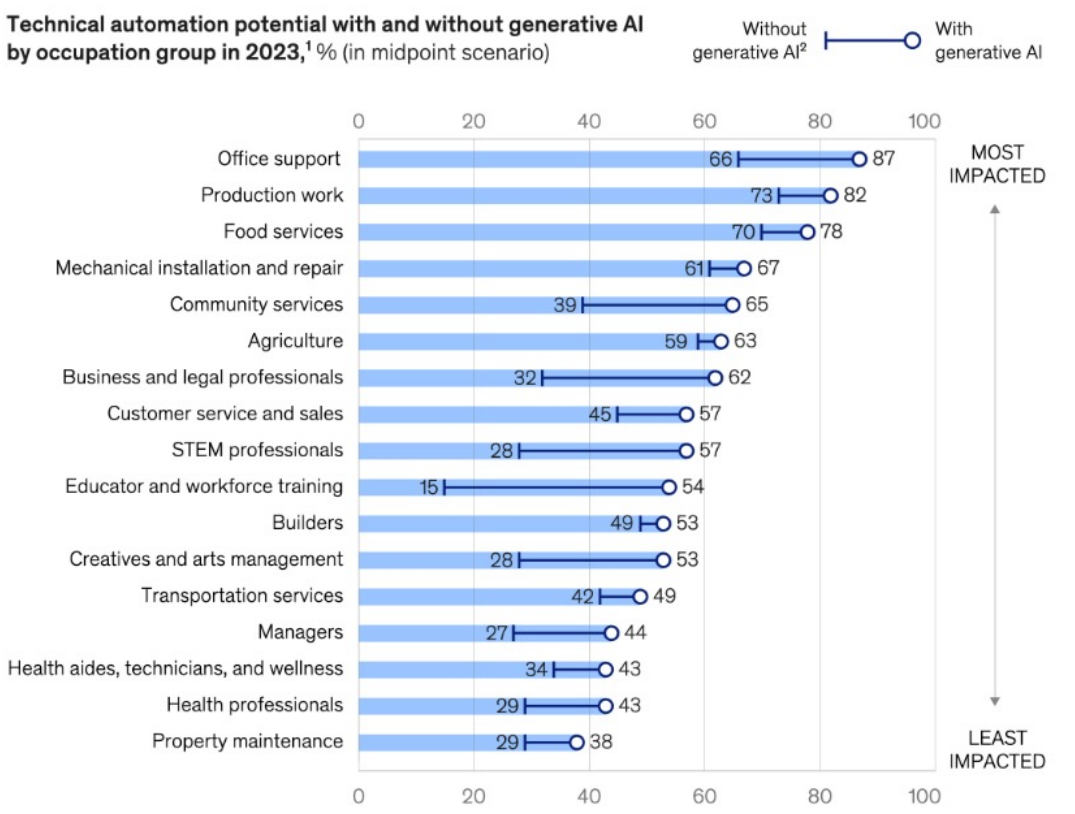
AI & Democracy

AI & Future of work

AI & Rights equality

AI & Medicine

Advances in AI's technical capabilities could have the most impact on activities performed by educators, professionals, and creatives.



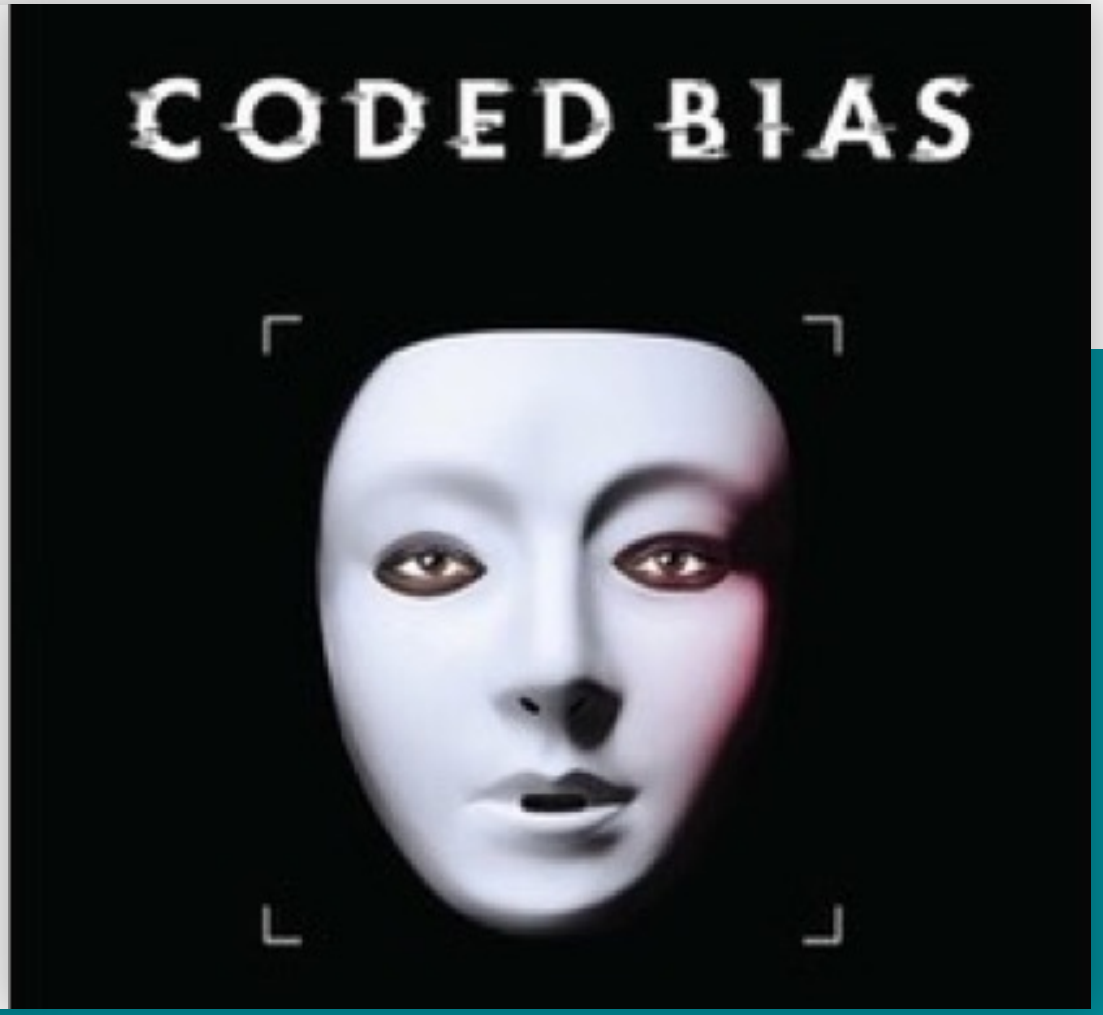
https://www.mckinsey.com/featured-insights/sustainable-inclusive-growth/chart-of-the-day/dont-ask-me-ai-just-work-here

AI's Impact in Society

AI can significantly impact society due to its large range of applications

Some dimensions

- **AI & Democracy**
- **AI & Future of work**
- **AI & Rights equality**
- **AI & Medicine**



https://www.netflix.com/title/81328723

AI's Impact in Society

AI can significantly impact society due to its large range of applications

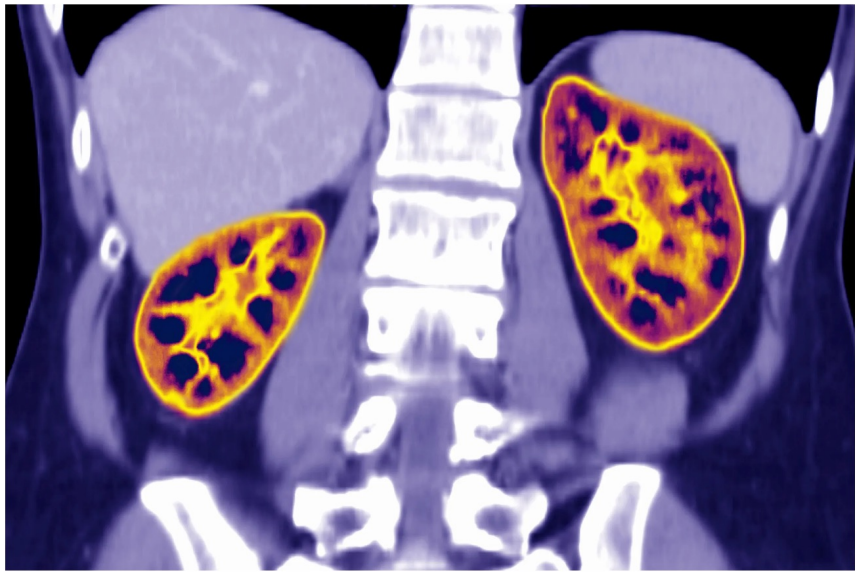
Some dimensions

- **AI & Democracy**
- **AI & Future of work**
- **AI & Rights equality**
- **AI & Medicine**

TOM SIMONITE BUSINESS OCT 26, 2020 7:00 AM

How an Algorithm Blocked Kidney Transplants to Black Patients

A formula for assessing the gravity of kidney disease is one of many that is adjusted for race. The practice can exacerbate health disparities.



A score known as eGFR aims to reflect the seriousness of a patient's kidney disease. PHOTOGRAPH: JAMES CAVALLINI/SCIENCE SOURCE

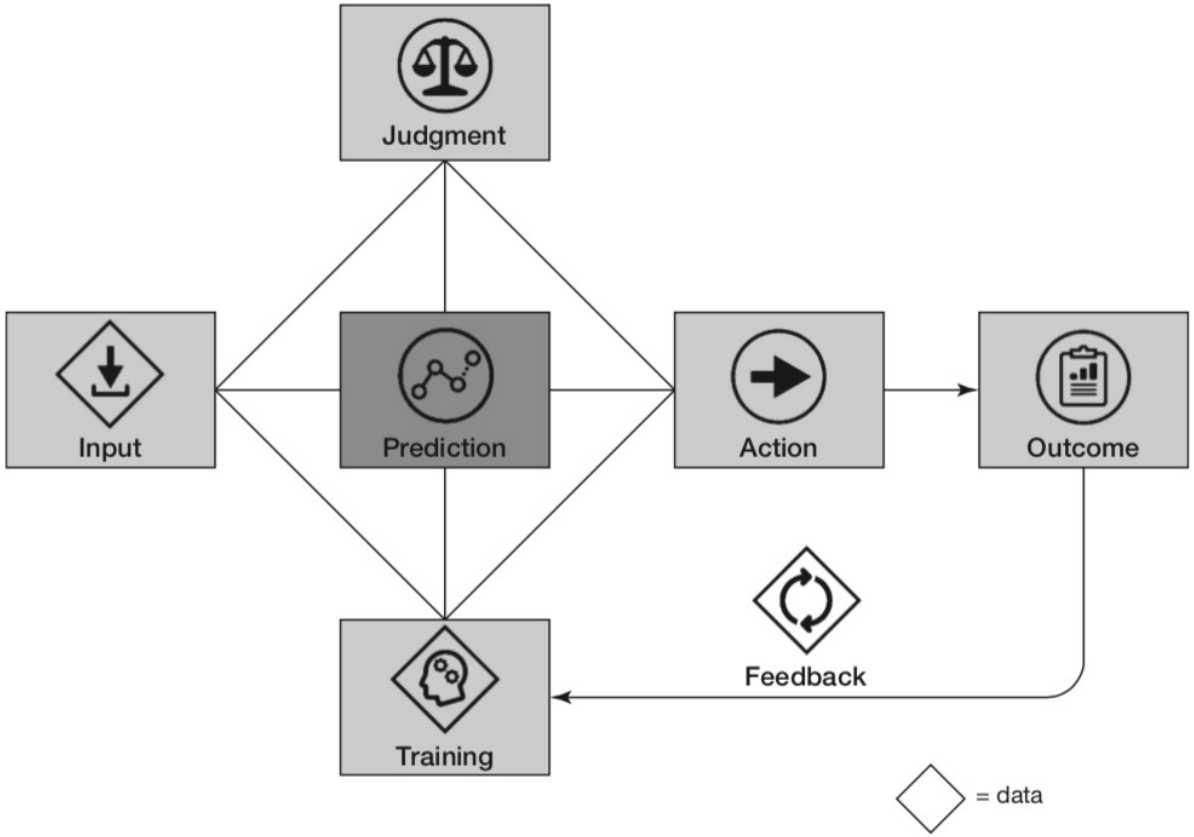
The AI Database →

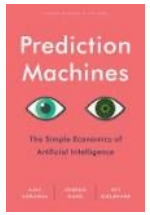
APPLICATION: RECOMMENDATION ALGORITHM, ETHICS

END USER: BIG COMPANY, SMALL COMPANY SECTOR: HEALTH CARE, RESEARCH

https://www.wired.com/story/how-algorithm-blocked-kidney-transplants-black-patients/

AI impact's society because we take decisions based on algorithm results… Let's have a look into the anatomy of a decision





https://www.predictionmachines.ai

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

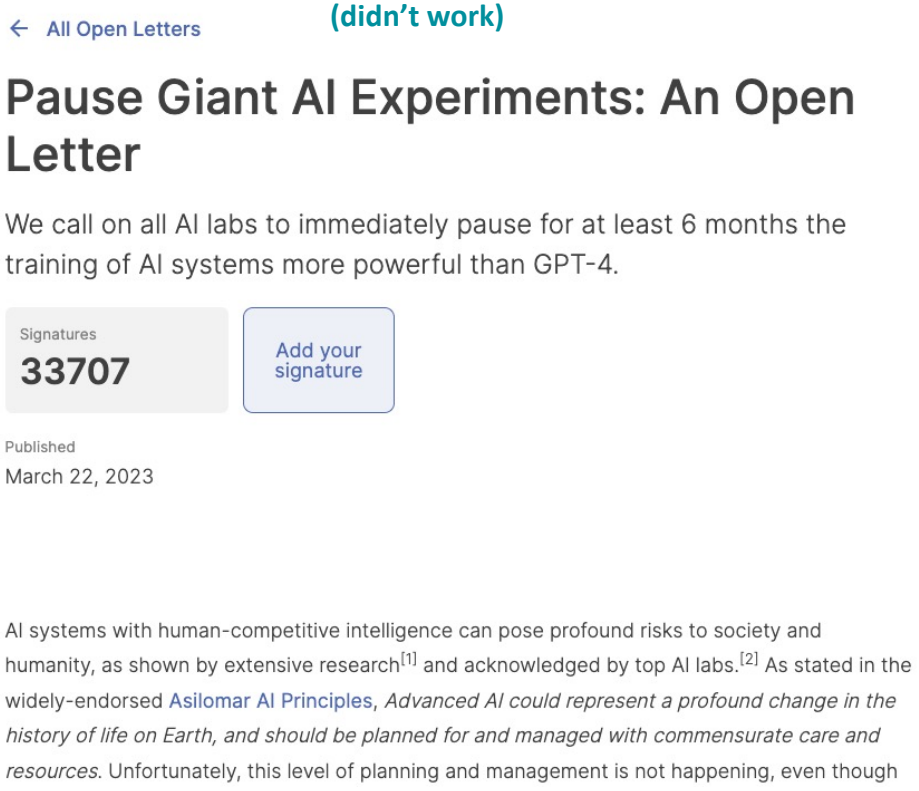
With the fast pace of technology and its expected impact in society what can we do?

HOSPITAL DA LUZ
LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

With the fast pace of technology and its expected impact in society what can we do?

Pause

(didn't work)



← All Open Letters

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
33707

Add your signature

Published
March 22, 2023

AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research[1] and acknowledged by top AI labs.[2] As stated in the widely-endorsed Asilomar AI Principles, *Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.* Unfortunately, this level of planning and management is not happening, even though

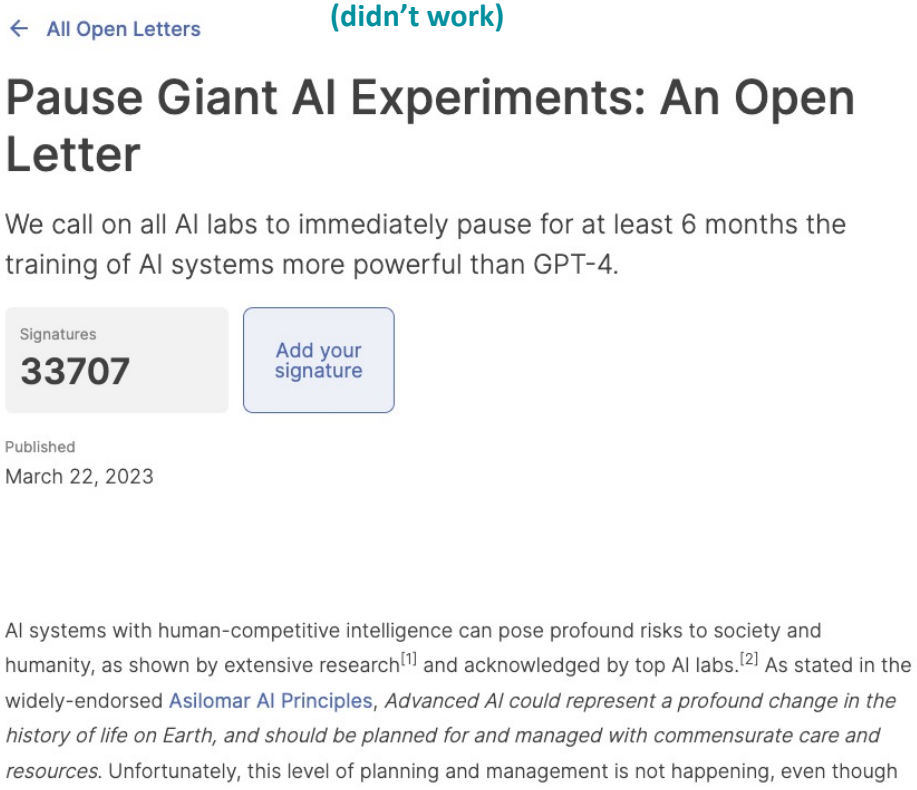
https://futureoflife.org/open-letter/pause-giant-ai-experiments/

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

With the fast pace of technology and its expected impact in society what can we do?

Pause

(didn't work)



← All Open Letters

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

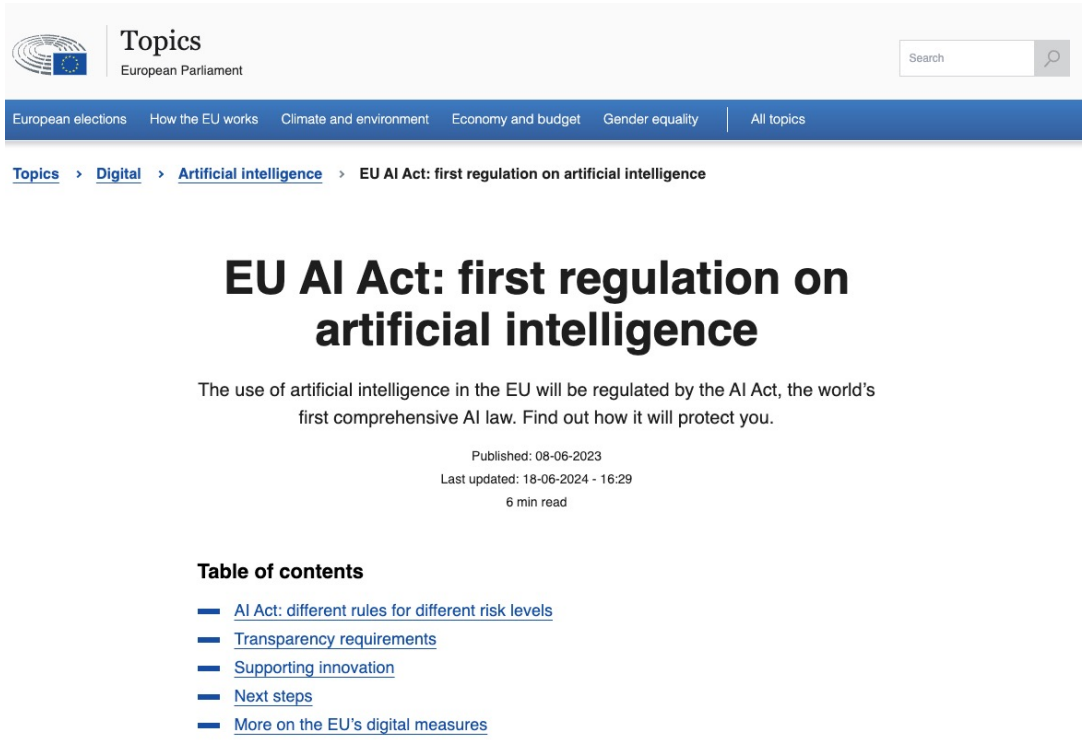
Signatures
33707

Add your signature

Published
March 22, 2023

AI systems with human-competitive intelligence can pose profound risks to society and humanity, as shown by extensive research[1] and acknowledged by top AI labs.[2] As stated in the widely-endorsed Asilomar AI Principles, *Advanced AI could represent a profound change in the history of life on Earth, and should be planned for and managed with commensurate care and resources.* Unfortunately, this level of planning and management is not happening, even though

Regulate



Topics
European Parliament

Search

European elections | How the EU works | Climate and environment | Economy and budget | Gender equality | All topics

Topics › Digital › Artificial intelligence › EU AI Act: first regulation on artificial intelligence

EU AI Act: first regulation on artificial intelligence

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.

Published: 08-06-2023
Last updated: 18-06-2024 - 16:29
6 min read

Table of contents

- AI Act: different rules for different risk levels
- Transparency requirements
- Supporting innovation
- Next steps
- More on the EU's digital measures

https://futureoflife.org/open-letter/pause-giant-ai-experiments/
https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Some highlights regarding the AI Act

What is AI?

Initially proposed by the EU commission [1]

- "'artificial intelligence system' (AI system) means **software** that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of **human-defined objectives**, generate outputs such as **content**, **predictions**, **recommendations**, or decisions influencing the environments they interact with"
  - Annex I (Machine learning, Logic and knowledge-based approaches and statistical approaches)

Amendments [2]

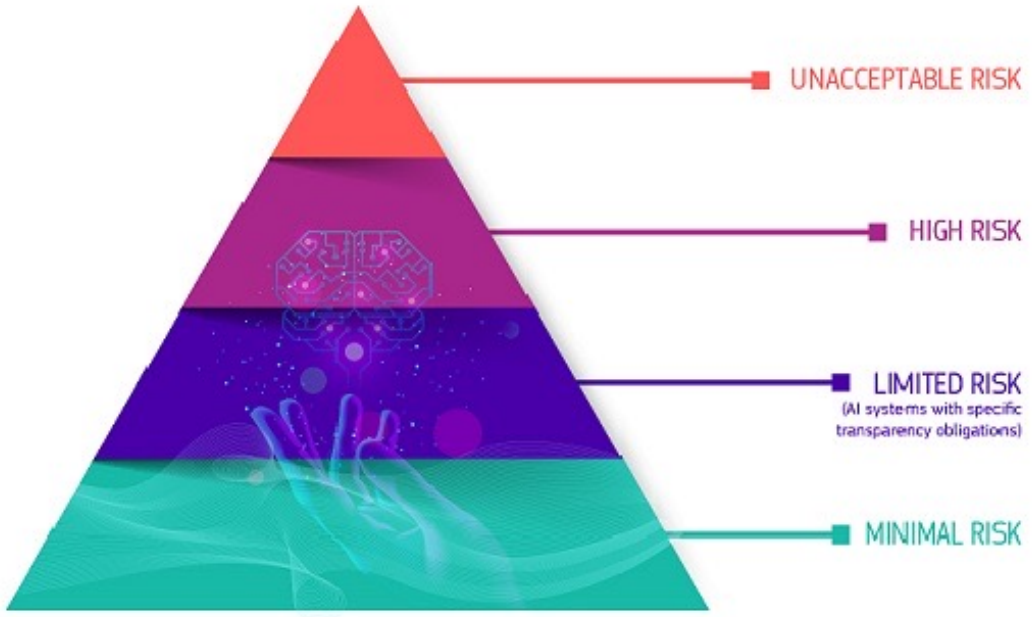
- "'artificial intelligence system' (AI system) means a **machine-based system** that is designed to operate with **varying levels of autonomy** and that can, for **explicit or implicit objectives**, generate outputs such as predictions, recommendations, or decisions, that influence physical or virtual environments;"

[1]EU Proposal - Laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending contain union legislative acts (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206)
[2]Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html)

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Some highlights regarding the AI Act

- Risk based approach based on the application and not on the technology itself



- There will be liabilities to companies that don't comply with AI Act

https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai

Why do we need responsible AI (in medicine)?

Why do we need responsible AI (in medicine)?

Use Case

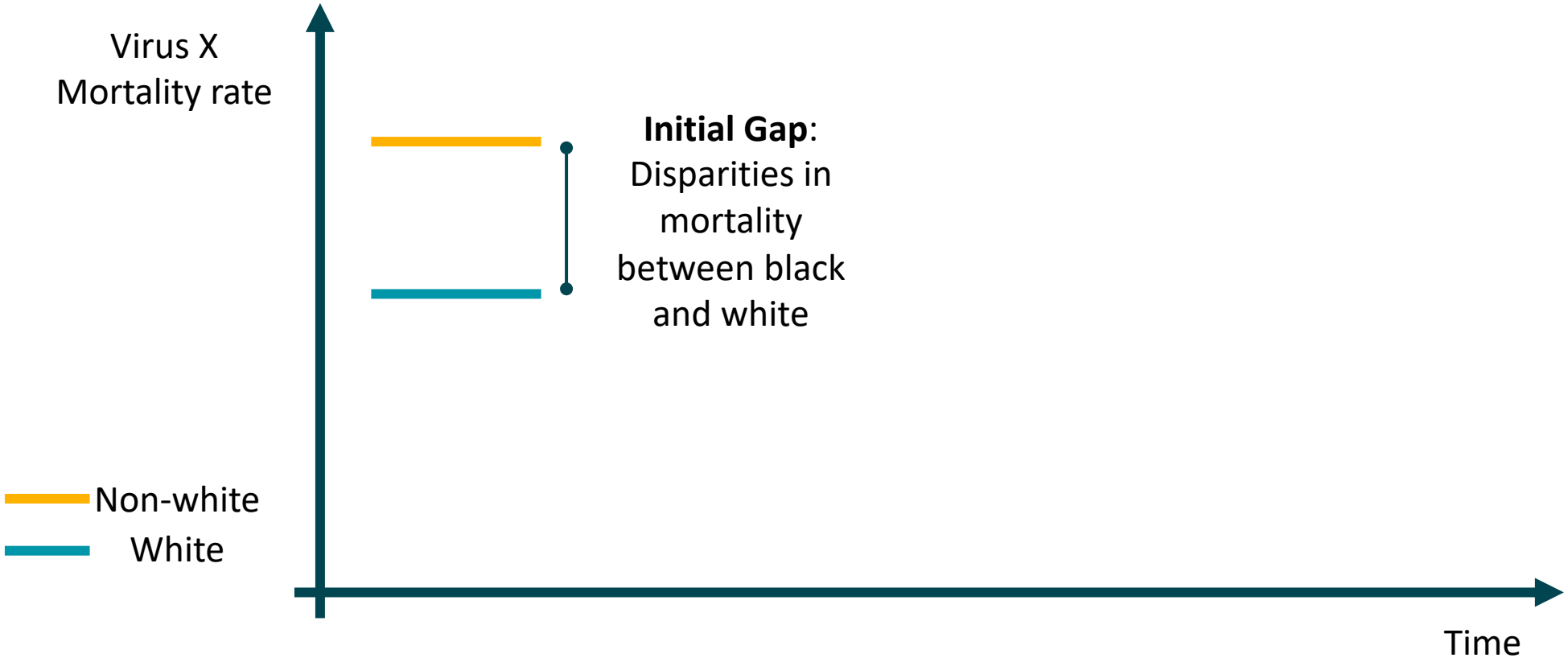
- Virus X has high mortality rate on people with certain medical condition
- A new drug is effective, but it is very expensive
- DGS decides to create a ML model using data from EHR and Social Media to predict people at risk to prioritize who must get the treatment

Based on Pedro Saleiro's Talk in LxMLS School 2021

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Why do we need responsible AI (in medicine)?

Use Case

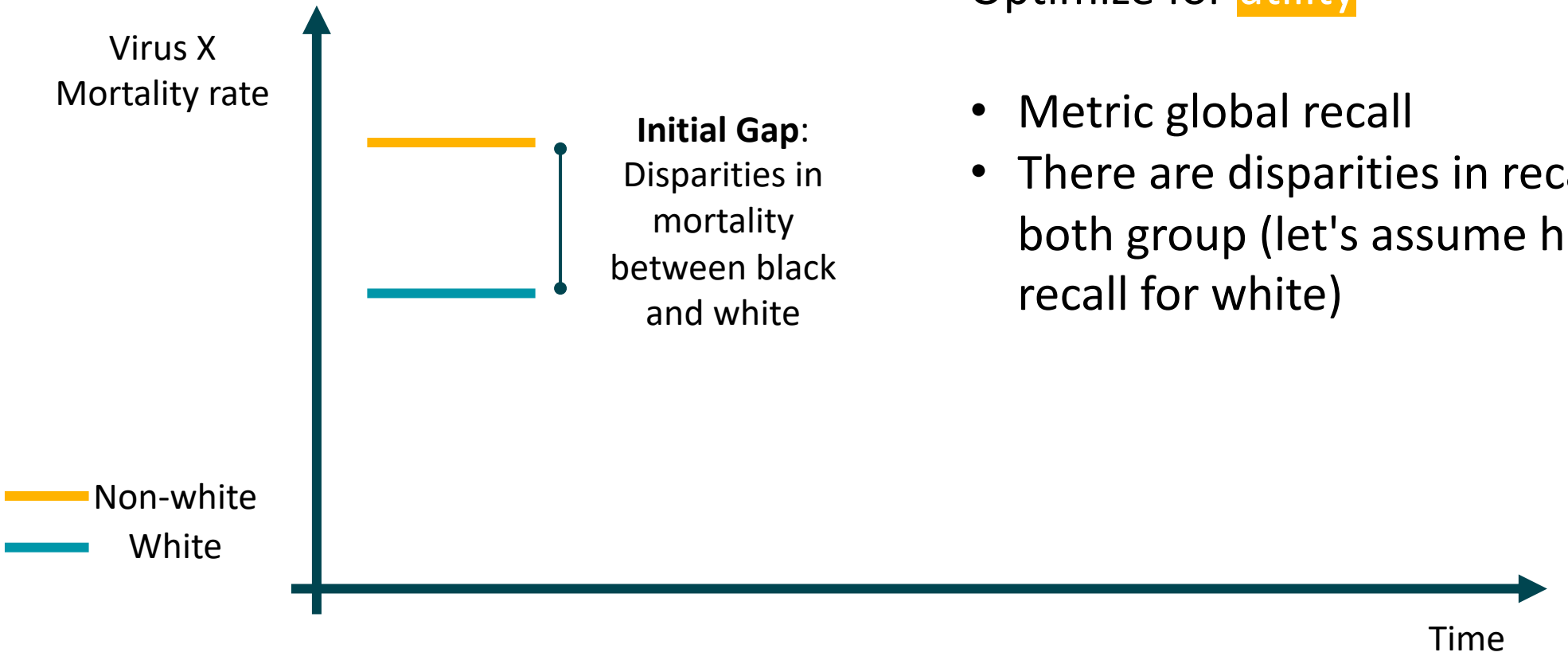


Based on Pedro Saleiro's Talk in LxMLS School 2021

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Why do we need responsible AI (in medicine)?

Use Case



Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

Optimize for utility

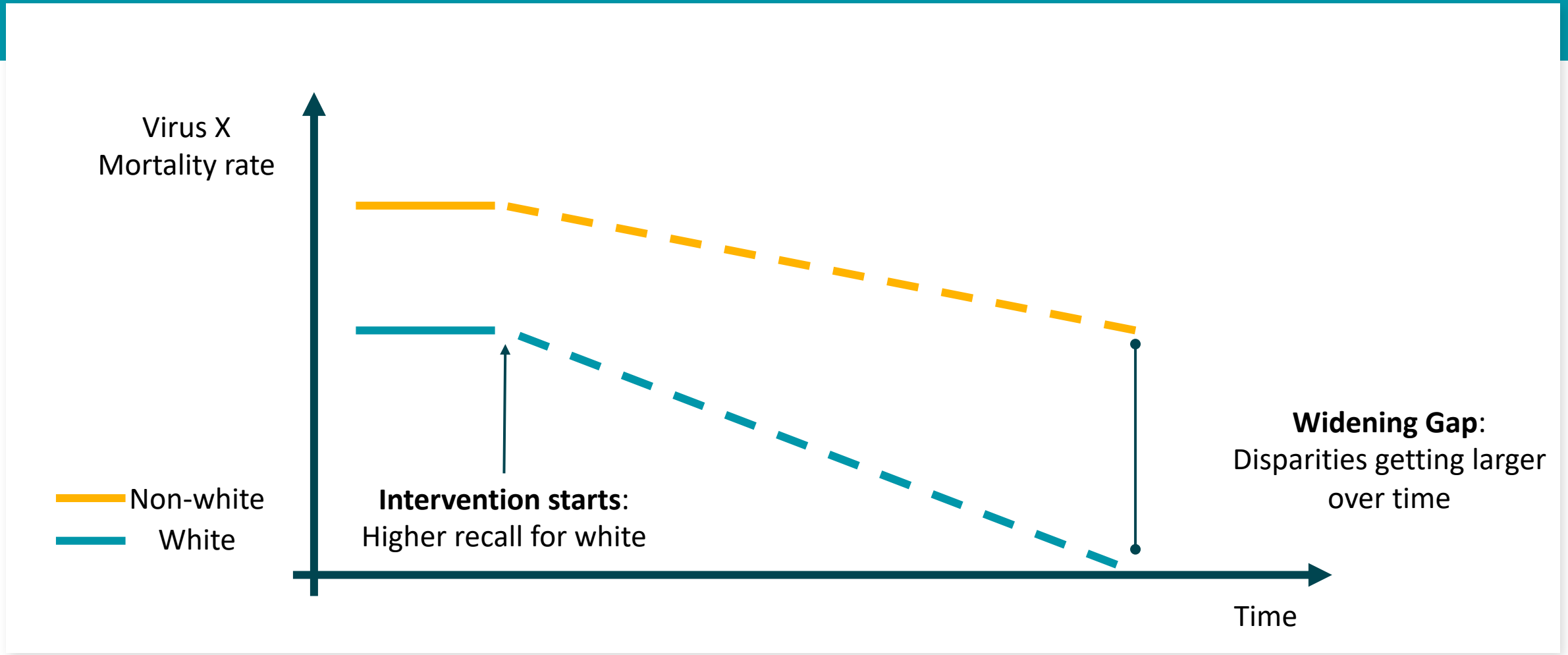
- Metric global recall
- There are disparities in recall for both group (let's assume higher recall for white)

Based on Pedro Saleiro's Talk in LxMLS School 2021

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Why do we need responsible AI (in medicine)?

Use Case



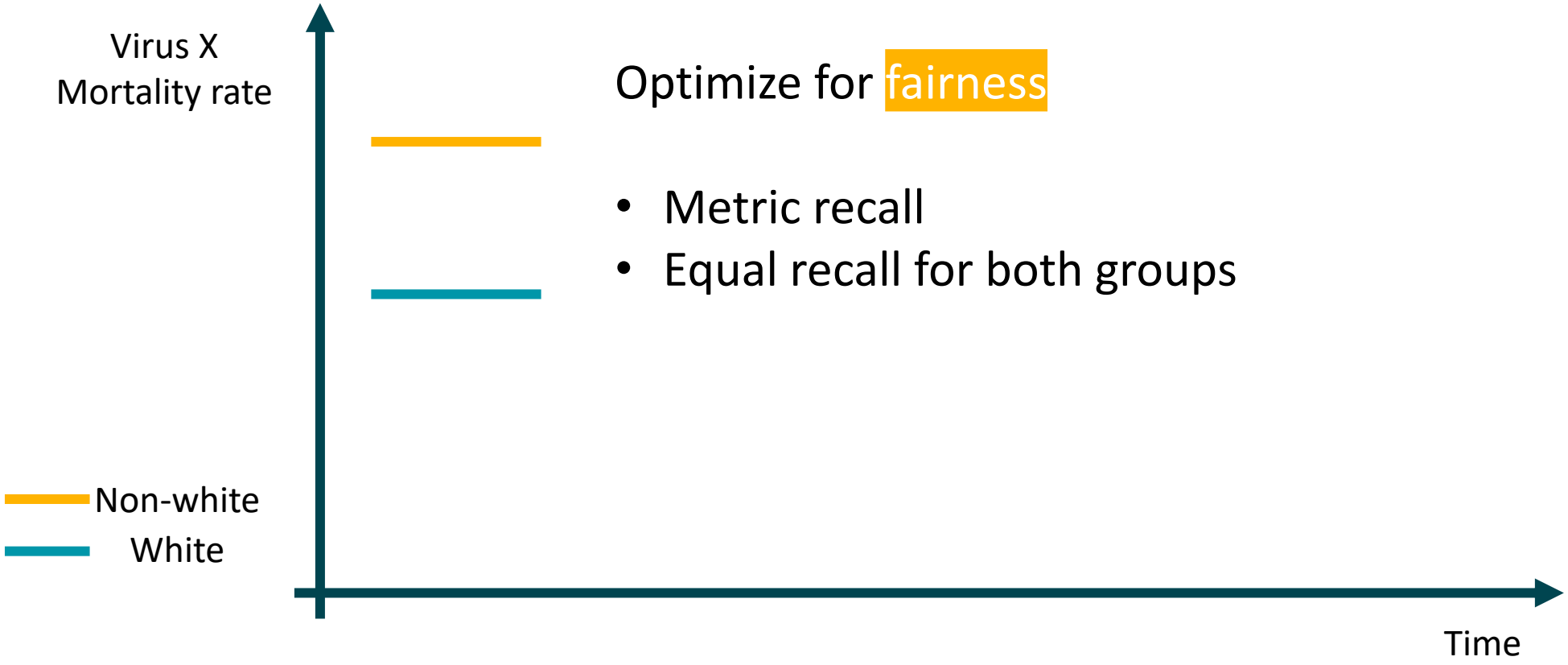
Illustrative example assuming intervention is effective

Based on Pedro Saleiro's Talk in LxMLS School 2021

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Why do we need responsible AI (in medicine)?

Use Case



Optimize for fairness

- Metric recall
- Equal recall for both groups

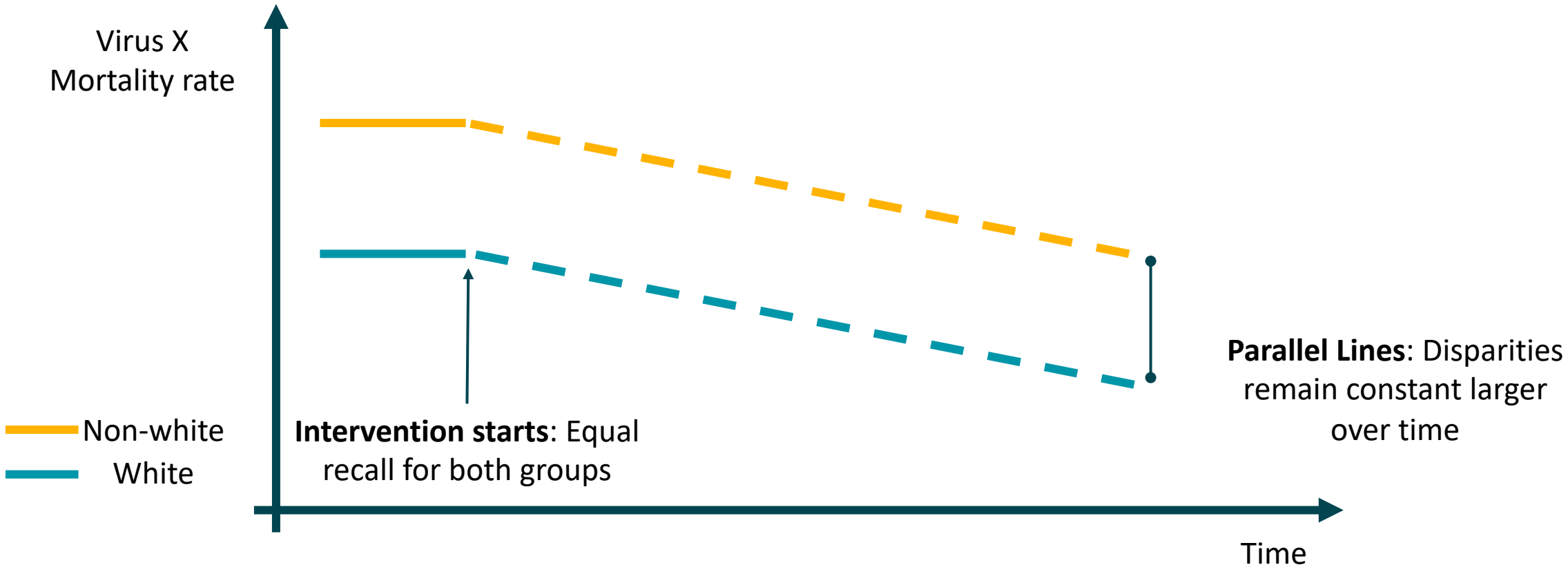
Illustrative example assuming intervention is effective

Based on Pedro Saleiro's Talk in LxMLS School 2021

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Why do we need responsible AI (in medicine)?

Use Case



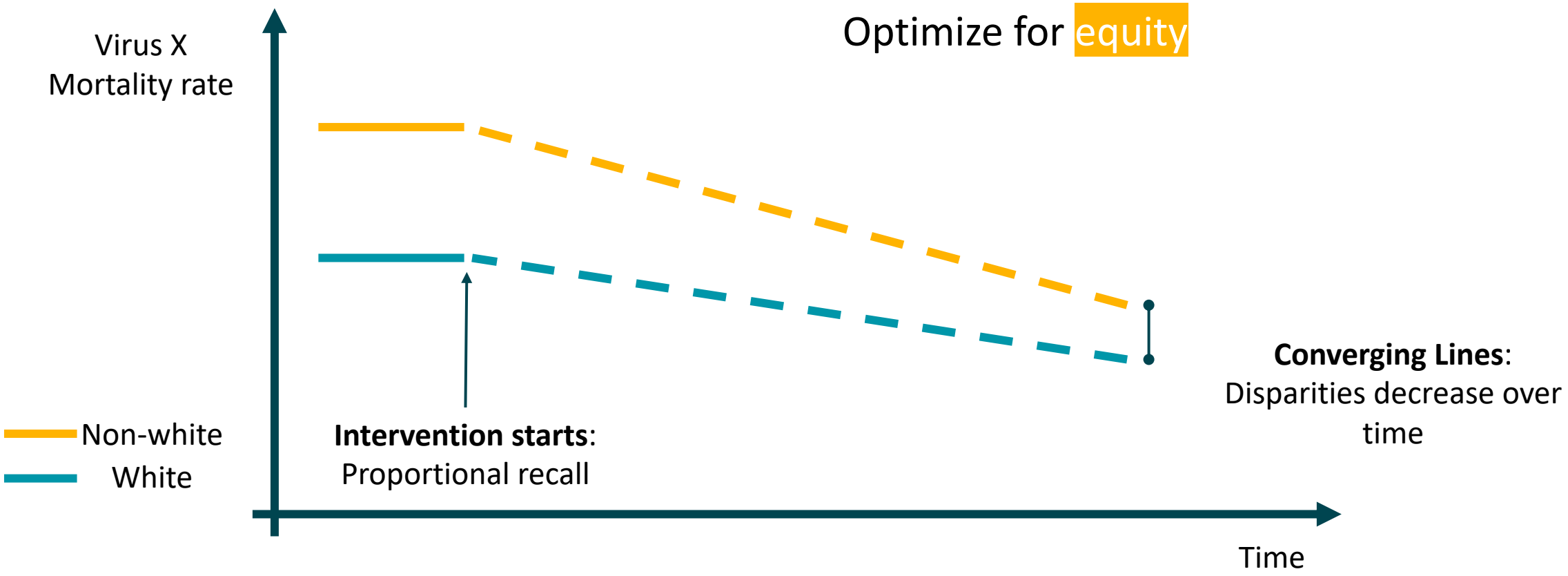
Illustrative example assuming intervention is effective

Based on Pedro Saleiro's Talk in LxMLS School 2021

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Why do we need responsible AI (in medicine)?

Use Case



Illustrative example assuming intervention is effective

Based on Pedro Saleiro's Talk in LxMLS School 2021

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Why do we need responsible AI (in medicine)?

Healthcare is used to build evidence to use new technologies

Comment

https://doi.org/10.1038/s41591-024-02853-7

To do no harm — and the most good — with AI in health care

Carey Beth Goldberg, Laura Adams, David Blumenthal, Patricia Flatley Brennan, Noah Brown, Atul J. Butte, Morgan Cheatham, Dave deBronkart, Jennifer Dixon, Jeffrey Drazen, Barbara J. Evans, Sara M. Hoffman, Chris Holmes, Peter Lee, Arjun Kumar Manrai, Gilbert S. Omenn, Jonathan B. Perlin, Rachel Ramoni, Guillermo Sapiro, Rupa Sarkar, Harpreet Sood, Effy Vayena, Isaac S. Kohane & the RAISE Consortium

Drawing from real-life scenarios and insights shared at the RAISE (Responsible AI for Social and Ethical Healthcare) conference, we highlight the critical need for AI in health care (AIH) to primarily benefit patients and address current shortcomings in health care systems such as medical errors and access disparities.

The conference, embodying a sense of responsibility and urgency, emphasized that AIH should enhance patient care, support health care professionals, and be accessible and safe for all. The discussions revolved around immediate actions for health care leaders, such as adopting AI to augment clinical practice, establishing transparent financial models, and guiding optimal AI use. The importance of AI as a complementary tool rather than as a replacement in health care, the harms to be minimized are those inflicted by the current health care system, from medical errors to lack of access.

Given such ongoing problems and recent technological leaps, adopting AIH where it could help is a matter of urgency. At a moment when the complexity of modern medicine has surpassed the capacity of the human mind, only AIH will be able to perform many tasks. AIH thus seems to offer unparalleled potential for further medical progress, including for precision medicine — the right therapy, for the right patient, at the right time. Thus, it is in the interest of the public and the medical profession to hasten its adoption, so long as it is used safely and made maximally accessible for all. It is also urgent to determine how AIH can best deliver tangible benefits, including how it can help improve health and save lives in ways that will not otherwise happen. The public should not only be aware of this quest but also participate in it.

This consensus emerged at the RAISE (Responsible AI for Social and Ethical Healthcare) conference, which was organized by the Department of Biomedical Informatics at Harvard Medical School. The conference was held October 29–31, 2023, in Cape Neddick, Maine. The

Goldberg, C.B., Adams, L., Blumenthal, D. *et al.* To do no harm — and the most good — with AI in health care. *Nat Med* **30**, 623–627 (2024). https://doi.org/10.1038/s41591-024-02853-7

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

What is responsible AI ?

A structured process to develop and commercialize AI products

- **Responsible AI** is about implementing **principles** and **best practices** for building safe, secure, ethical and trustworthy **AI** systems.

What is responsible AI ?

There are multiple frameworks to implement responsible AI and manage risk

Some examples

NIST

Search NIST | Menu

Information Technology Laboratory

AI RISK MANAGEMENT FRAMEWORK

- AI RMF Development
- NIST AI RMF Playbook
- Engage
- Workshops & Events
- Related NIST Efforts
- Resources
- Perspectives
- FAQs
- AI @ NIST

On April 29, 2024, NIST released a draft publication based on the AI Risk Management Framework (AI RMF) to help manage the risk of Generative AI. The draft AI RMF Generative AI Profile can help organizations identify unique risks posed by generative AI and proposes actions for generative AI risk management that best aligns with their goals and priorities. Developed over the past year and drawing on input from the NIST generative AI public working group of more than 2,500 members, the guidance centers on a list of 12 risks and more than 400 actions that developers can take to manage them. More information is available here.

On April 30, 2024, NIST posted a crosswalk between the NIST AI Risk Management Framework (AI RMF) and the Japan AI Guidelines for Business (AI GfB.)

In collaboration with the private and public sectors, NIST has developed a framework to better manage risks to individuals, organizations, and society associated with artificial intelligence (AI). The NIST AI Risk Management Framework (AI RMF) is intended for voluntary use and to improve the ability to incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems.

Released on January 26, 2023, the Framework was developed through a consensus-driven, open, transparent, and collaborative process that included a Request for Information, several draft versions for public comments, multiple workshops, and other opportunities to provide input. It is intended to build on, align with, and support AI risk management efforts by others.

World Health Organization | Health Topics | Countries | Newsroom | Emergencies

WHO releases AI ethics and governance guidance for large multi-modal models

18 January 2024 | News release |Reading time: 3 min (841 words)

The World Health Organization (WHO) is releasing new guidance on the ethics and governance of large multi-modal models (LMMs) – a type of fast growing generative artificial intelligence (AI) technology with applications across health care.

https://www.who.int/news/item/18-01-2024-who-releases-ai-ethics-and-governance-guidance-for-large-multi-modal-models
https://www.nist.gov/itl/ai-risk-management-framework

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Key concepts in Responsible AI

Every technological company also has its own framework, but the key concepts are the same between them

- **Accountability**
- **Transparency and explainability**
- **Fairness**
- **Inclusiveness**
- **Privacy and security**

Key concepts in Responsible AI

Every technological company also has its own framework, but the key concepts are the same between them

- **Accountability**
- **Transparency and explainability**
- **Fairness**
- **Inclusiveness**
- **Privacy and security**

Transparency & Explainability

This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

Making outputs of AI interpretable

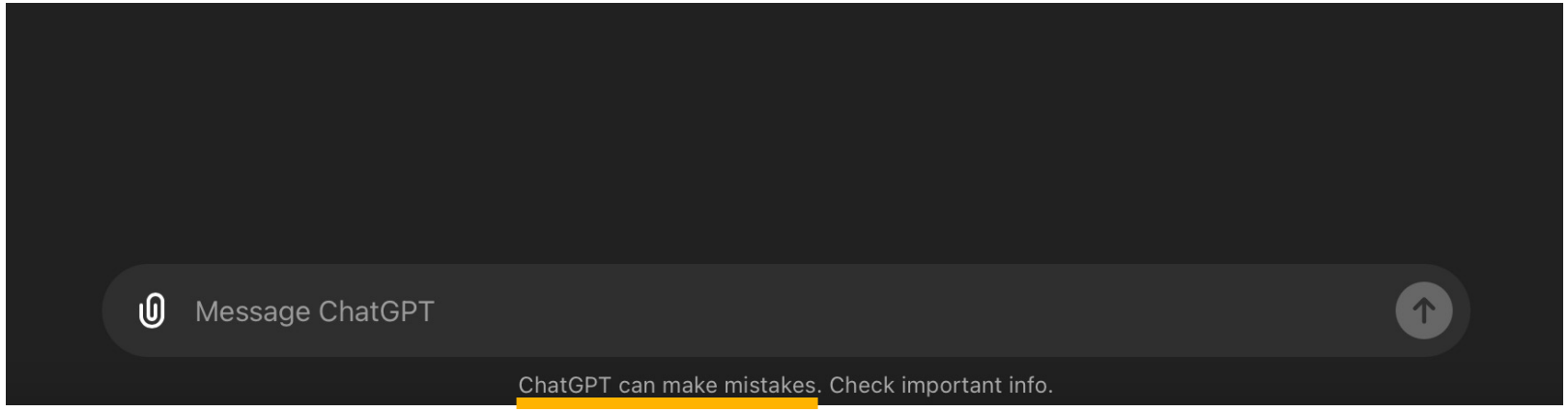
Transparency & Explainability

This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

- *People should know when they are interacting with a content generated by AI*
- *Companies should publish summaries of copyrighted data used for training (what data was used in developing the algorithm?)*

Making outputs of AI interpretable



HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Transparency & Explainability

This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

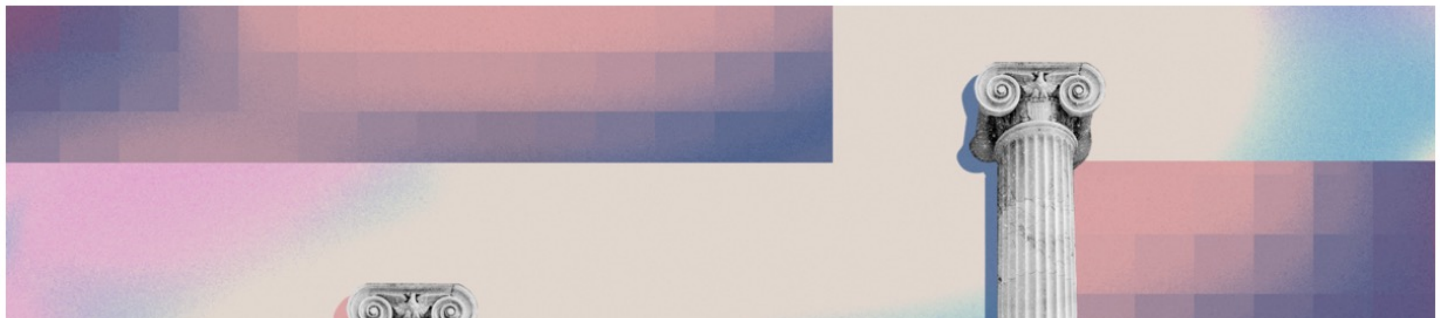
Making outputs of AI interpretable

Intellectual Property

Generative AI Has an Intellectual Property Problem

by Gil Appel, Juliana Neelbauer, and David A. Schweidel

April 07, 2023



https://hbr.org/2023/04/generative-ai-has-an-intellectual-property-problem

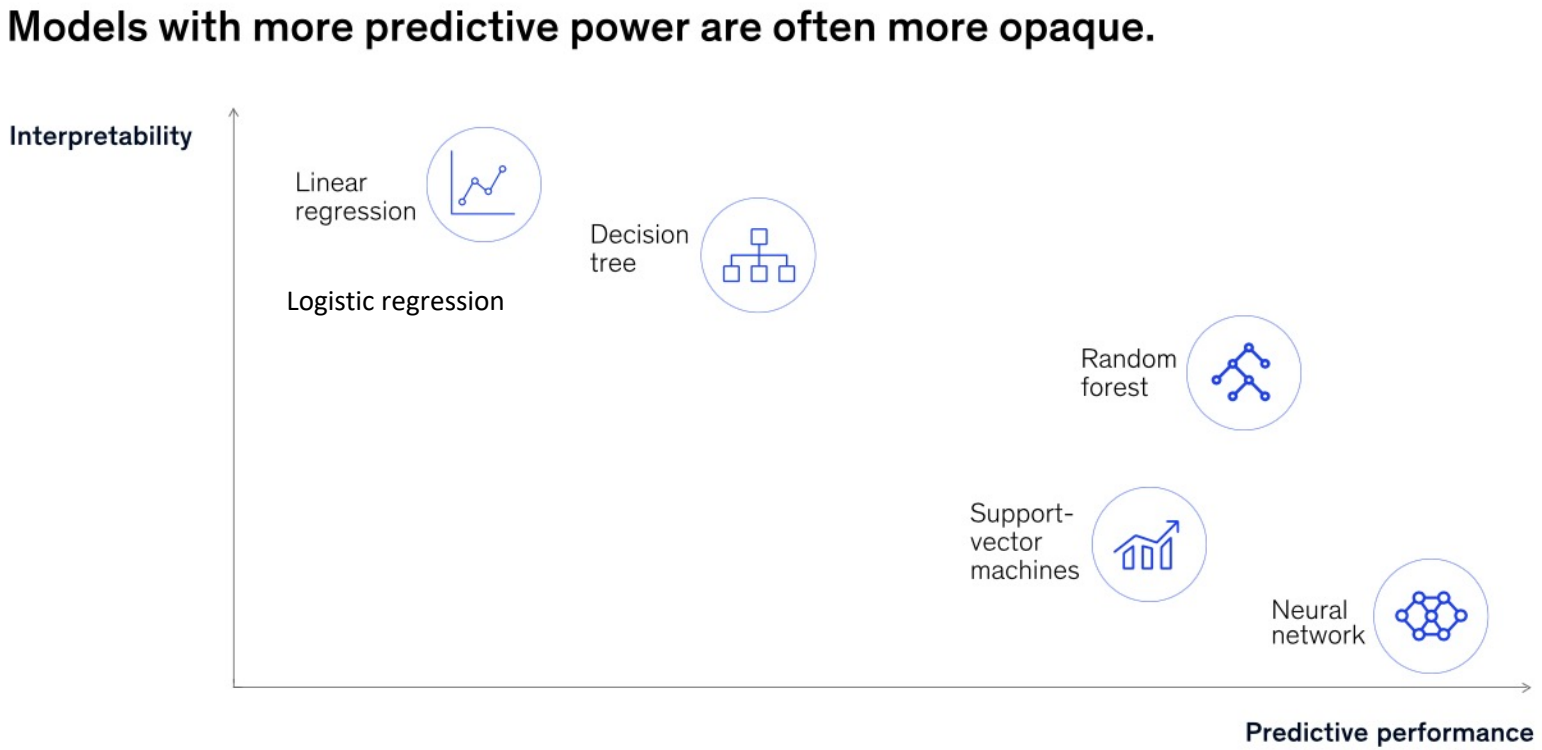
HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Transparency & Explainability

This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

Making outputs of AI interpretable

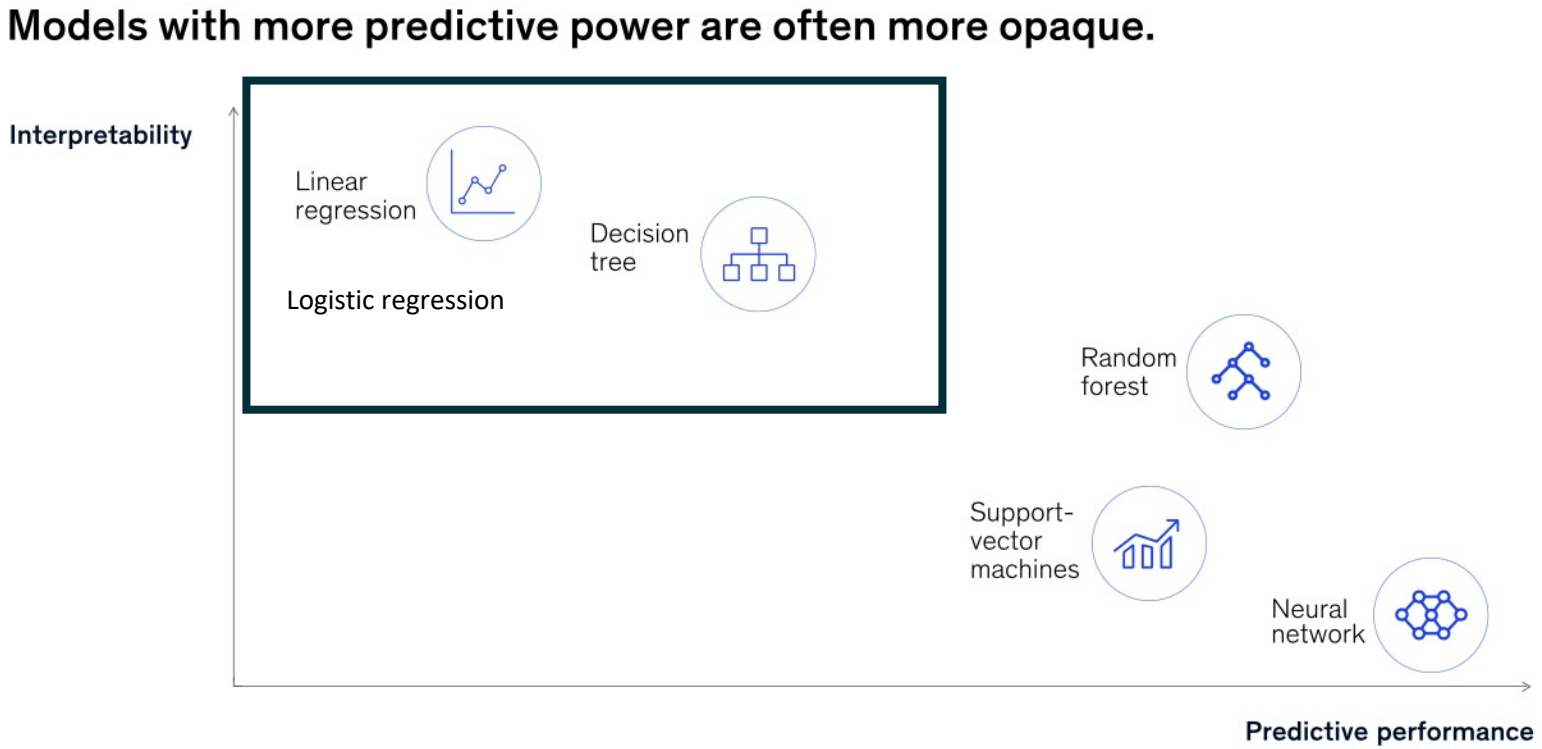


Transparency & Explainability

This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

Making outputs of AI interpretable

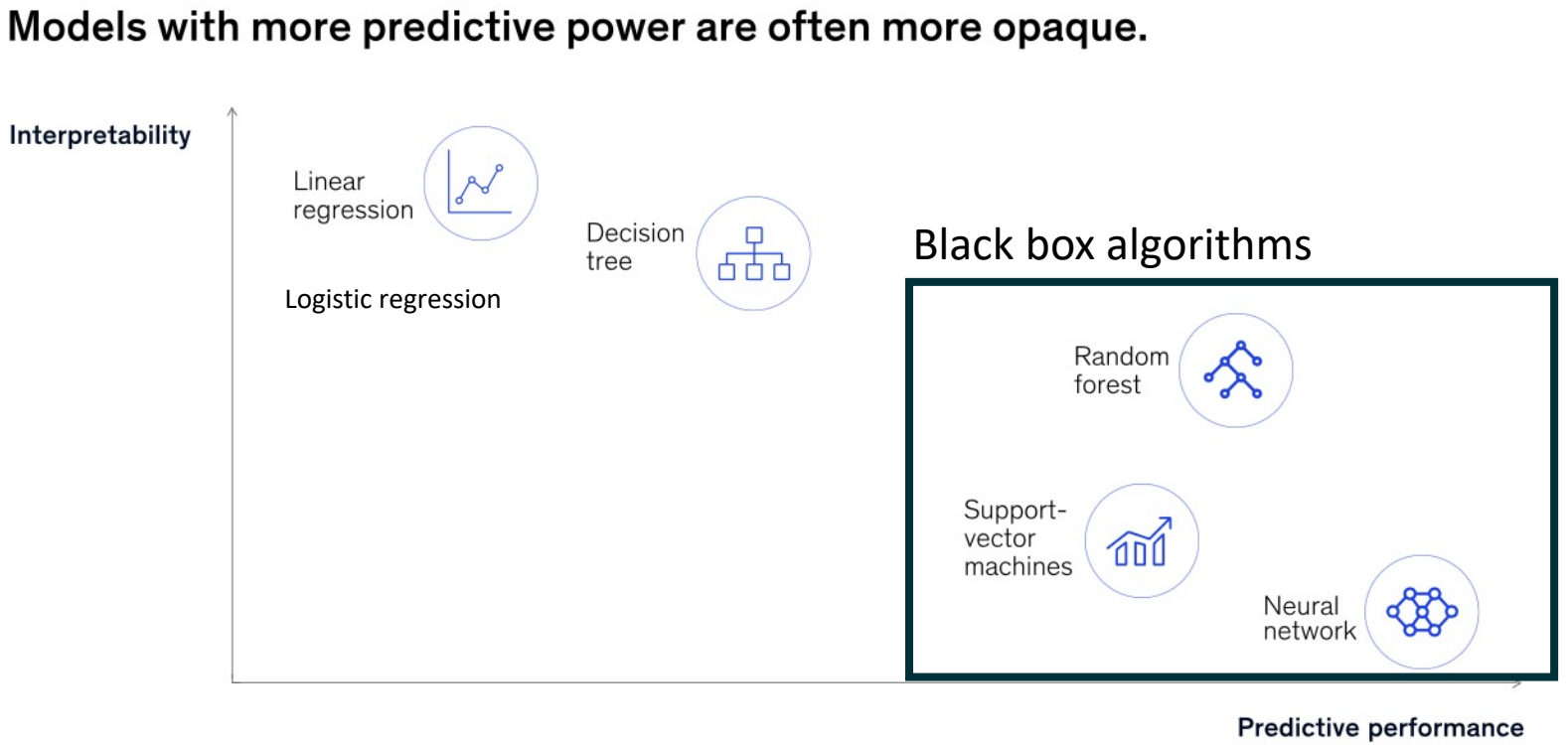


Transparency & Explainability

This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

Making outputs of AI interpretable



Transparency & Explainability

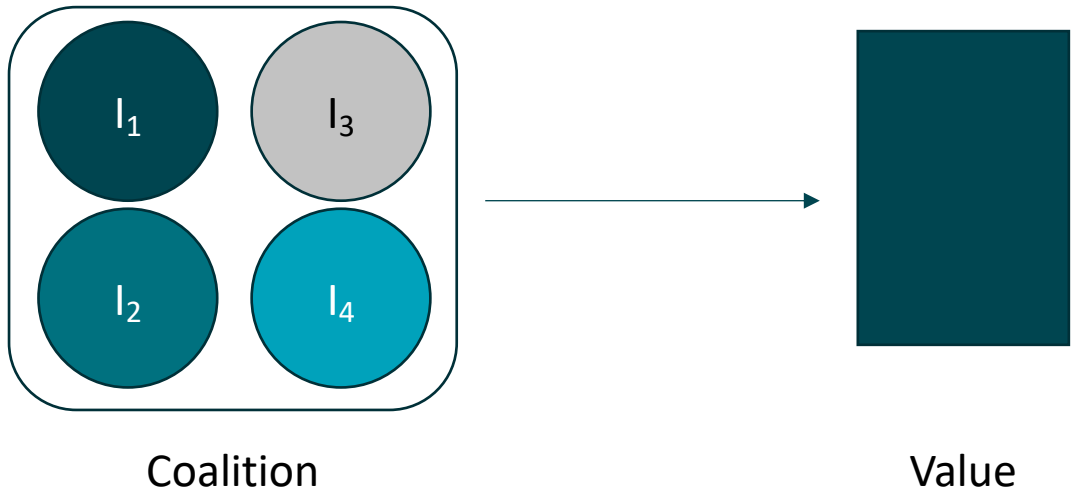
This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

Making outputs of AI interpretable

Shapley values

- Inspired in the game theory



- What is the contribution of everyone to the value?

Detailed video https://www.youtube.com/watch?v=VB9uV-x0gtg

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Transparency & Explainability

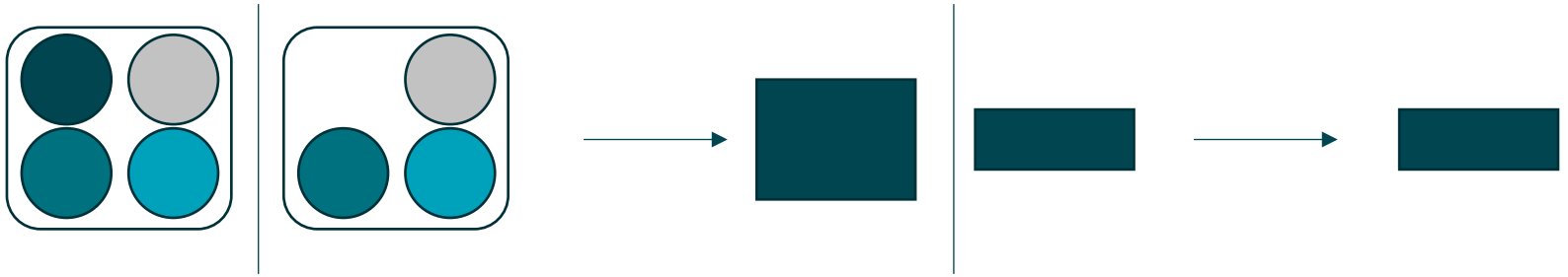
This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

Making outputs of AI interpretable

Shapley values

- Evaluate the value with and without a member of the coalition for all possible combinations



- Calculate the difference between scenarios

Detailed video https://www.youtube.com/watch?v=VB9uV-x0gtg

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Transparency & Explainability

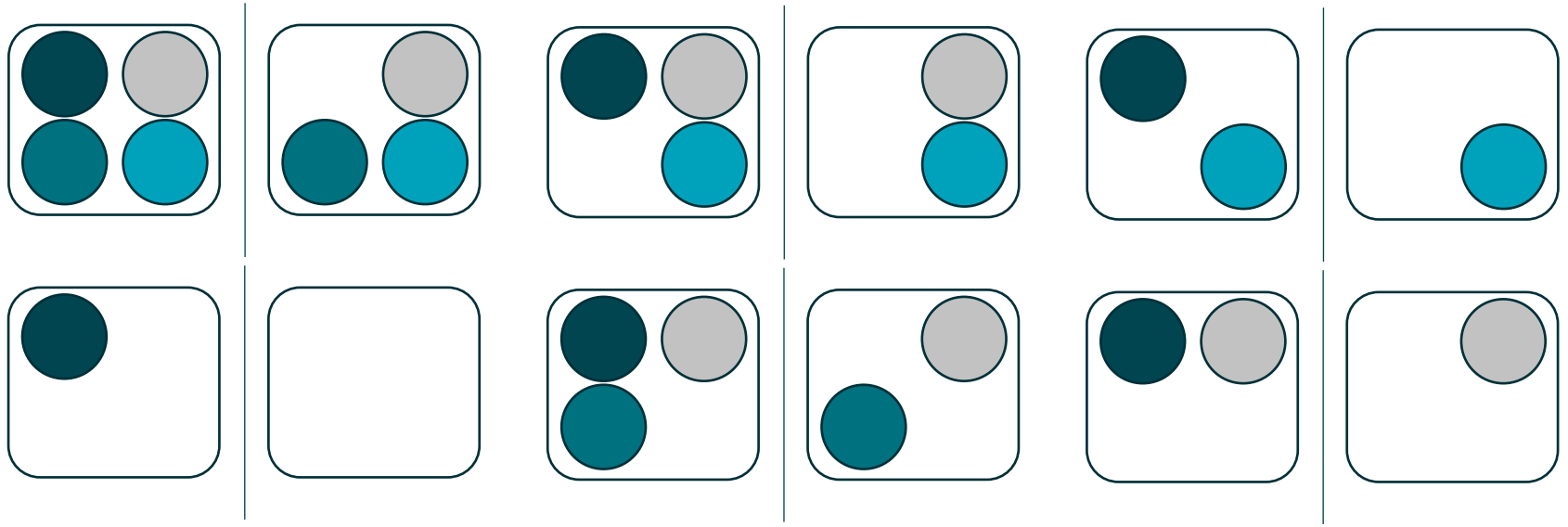
This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

Making outputs of AI interpretable

Shapley values

- Evaluate the value with and without a member of the coalition for all possible combinations



- Calculate the difference between scenarios

Detailed video https://www.youtube.com/watch?v=VB9uV-x0gtg

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Transparency & Explainability

This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

Making outputs of AI interpretable

Shapley values

- Evaluate the value with and without a member of the coalition for all possible combinations

- The average value of all scenarios is the Shapley value for this member.

Detailed video https://www.youtube.com/watch?v=VB9uV-x0gtg

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Transparency & Explainability

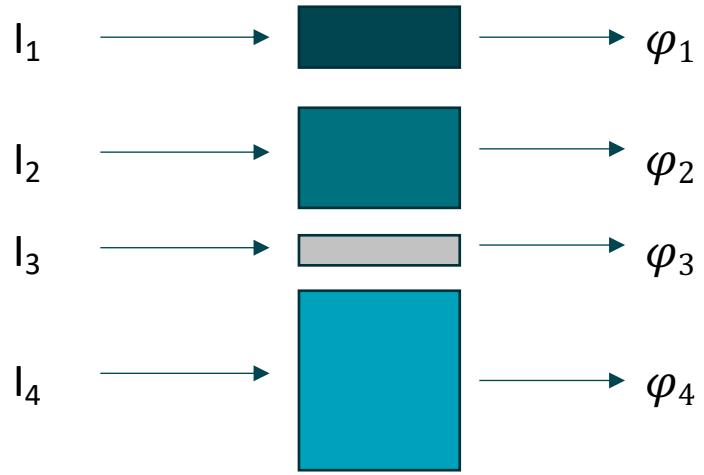
This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

Making outputs of AI interpretable

Shapley values

- Evaluate the value with and without a member of the coalition for all possible combinations



- The average value of all scenarios is the Shapley value for this member.
- The process is repeated for each member

Detailed video https://www.youtube.com/watch?v=VB9uV-x0gtg

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Transparency & Explainability

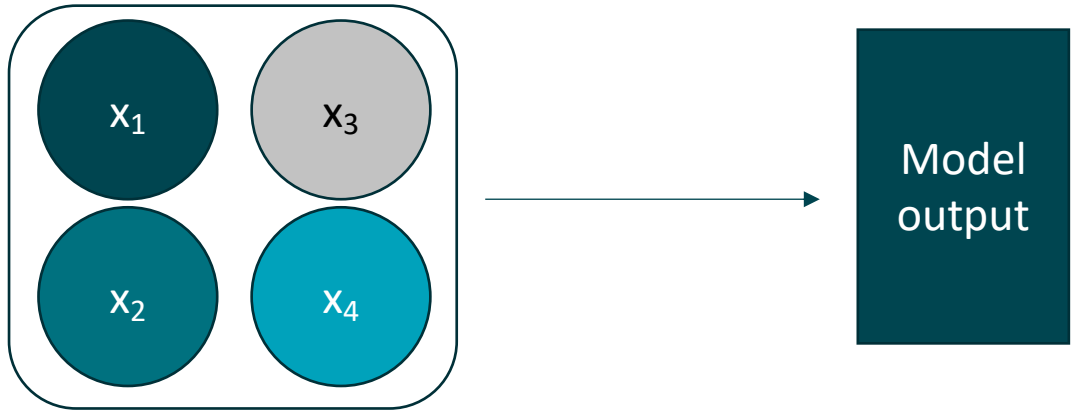
This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

Making outputs of AI interpretable

Shapley values and expandability (SHAP[1])

- The member of the coalition are the variables, and the value is the model output



Detailed video https://www.youtube.com/watch?v=VB9uV-x0gtg

[1]Original paper https://doi.org/10.48550/arXiv.1705.07874

Transparency & Explainability

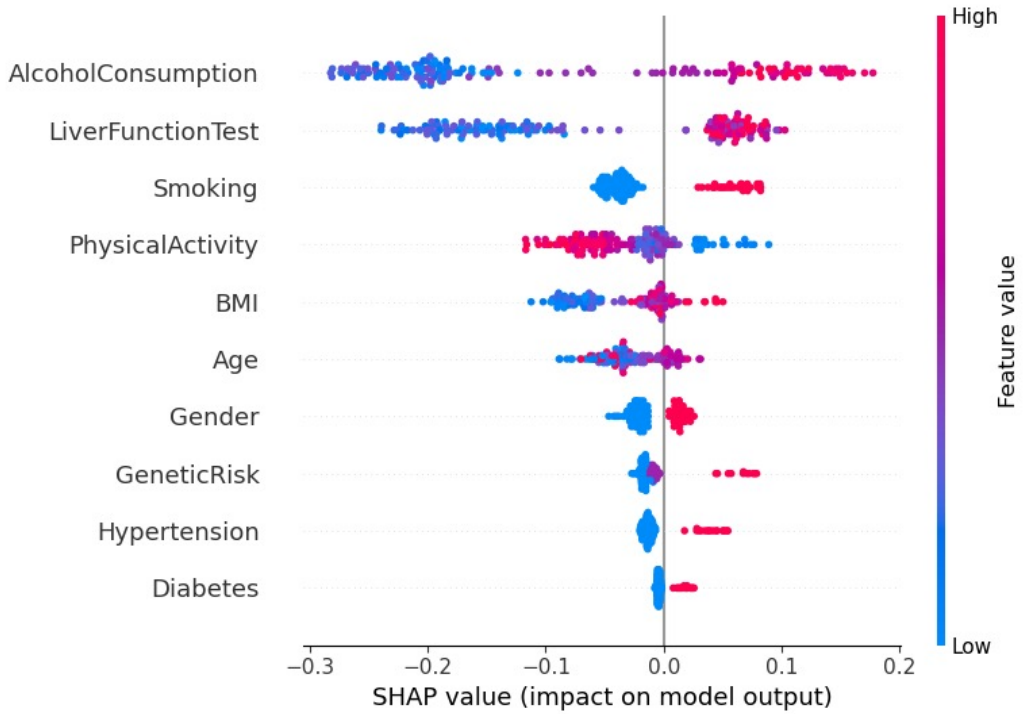
This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

Making outputs of AI interpretable

Shapley values and expandability (SHAP[1])

- The member of the coalition are the variables, and the value is the model output



Detailed video https://www.youtube.com/watch?v=VB9uV-x0gtg

[1]Original paper https://doi.org/10.48550/arXiv.1705.07874

Data: https://www.kaggle.com/datasets/rabieelkharoua/predict-liver-disease-1700-records-dataset

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

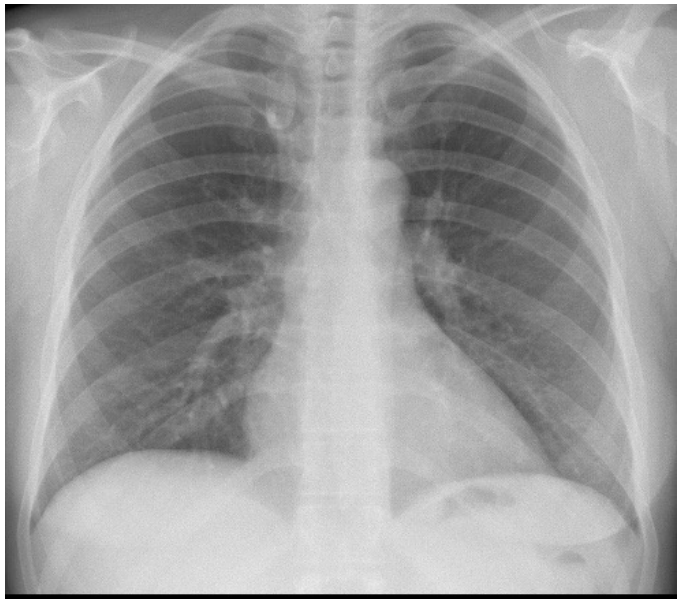

CODE
MEDICINE
CODE
MACHINE
LEARNING

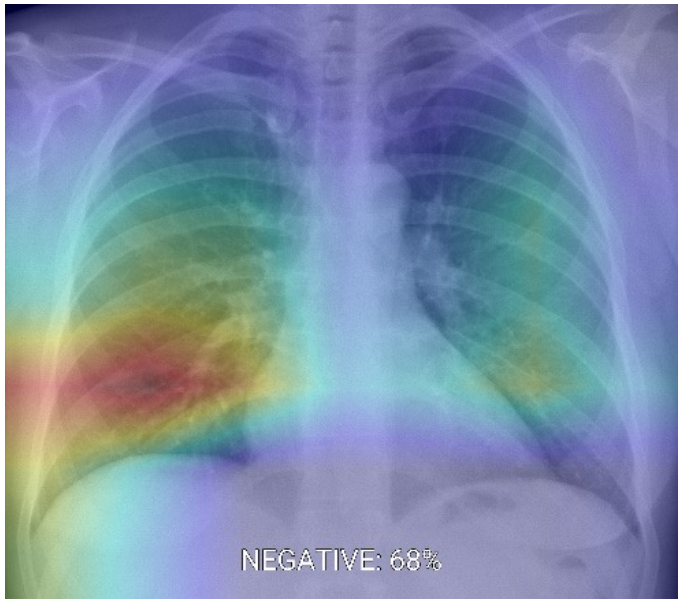
Transparency & Explainability

This principle is about transparency and responsible disclosure around AI systems to ensure that people understand when they are engaging with them and can challenge outcomes.

Disclose when AI is being used

Making outputs of AI interpretable





https://poloclub.github.io/cnn-explainer/

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Recap

Let's refresh some concepts

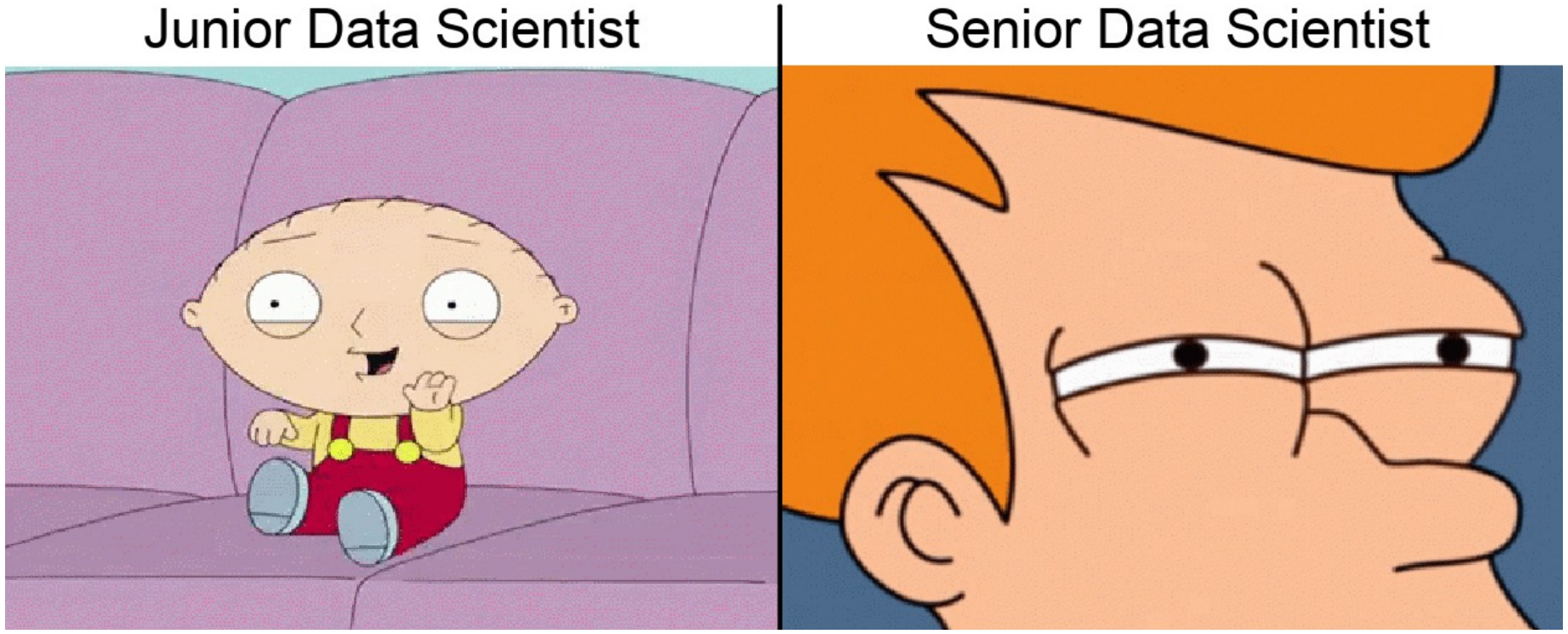
ML model returns above 99% accuracy on real-world data

Is this a good model?

Recap

Let's refresh some concepts

ML model returns above 99% accuracy on real-world data



Recap

Let's refresh some concepts

Confusion Matrix

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Calculate metrics →

https://en.wikipedia.org/wiki/Confusion_matrix

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Recap

Let's refresh some concepts

Confusion Matrix

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Calculate metrics →

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Error Rate

$$\text{Error Rate} = \frac{FP + FN}{TP + TN + FP + FN}$$

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 - score

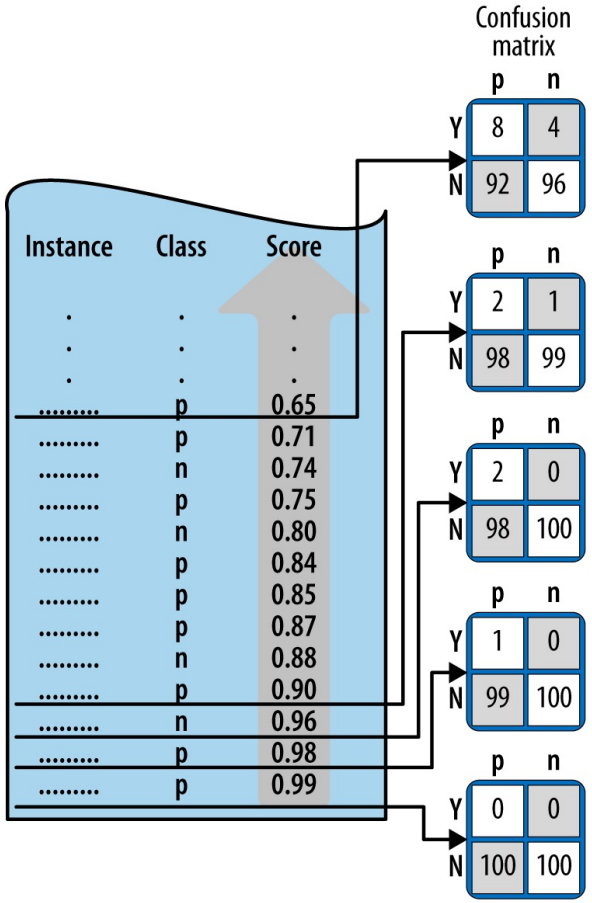
$$\text{F1 - score} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

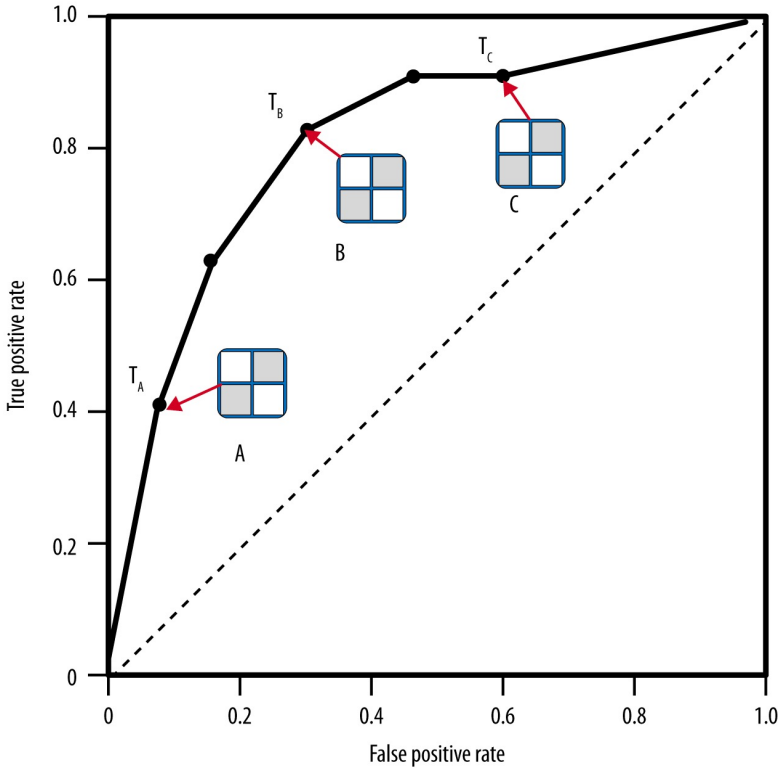
https://en.wikipedia.org/wiki/Confusion_matrix

Recap

Let's refresh some concepts

ROC Curve





HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Fairness

Receiver operating curve

- Predict liver disease diagnostics

Index	Score	Label_value
1	0.999800	1
2	0.986034	1
3	0.965328	1
4	0.927179	1
5	0.889940	0
6	0.837430	1
7	0.773571	1
8	0.690866	1
9	0.590255	1
10	0.468367	0
11	0.357743	1
12	0.276027	0
13	0.193268	0
14	0.145628	1
15	0.103324	0
16	0.062255	0
17	0.028513	0

Label: 1 positive, 0 negative

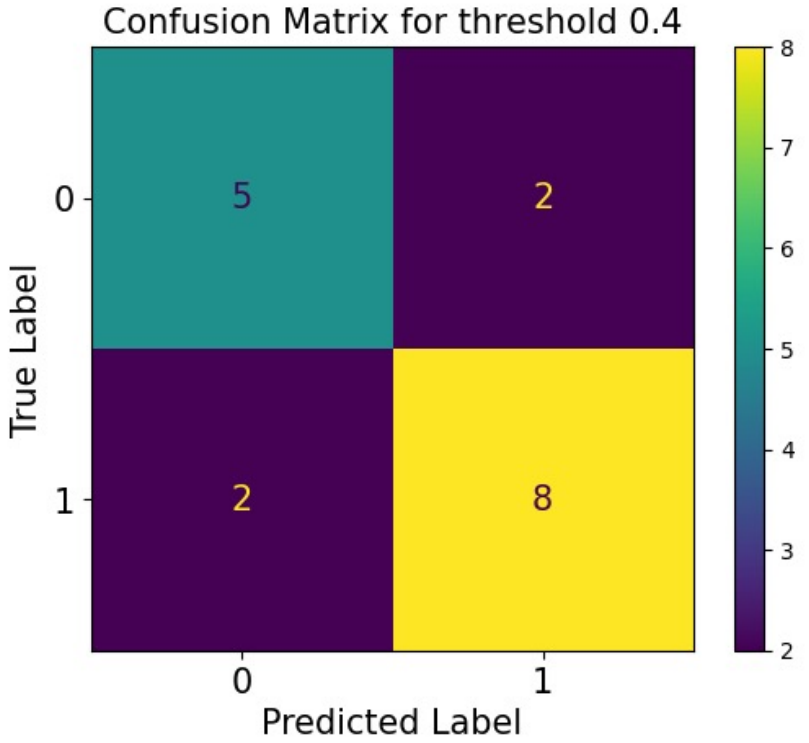
Fairness

Receiver operating curve

- Predict liver disease diagnostics

Index	Score	Label_value
1	0.999800	1
2	0.986034	1
3	0.965328	1
4	0.927179	1
5	0.889940	0
6	0.837430	1
7	0.773571	1
8	0.690866	1
9	0.590255	1
10	0.468367	0
11	0.357743	1
12	0.276027	0
13	0.193268	0
14	0.145628	1
15	0.103324	0
16	0.062255	0
17	0.028513	0

Label: 1 positive, 0 negative



TPR = tp / (tp + fn) = 8/10 = 0.8
FPR = fp / (fp +tn) = 2/7 = 0.3
Accuracy = 13/17 = 0.8

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Fairness

Receiver operating curve

- Predict liver disease diagnostics

Index	Score	Label_value	Race
1	0.999800	1	White
2	0.986034	1	Non-White
3	0.965328	1	Non-White
4	0.927179	1	White
5	0.889940	0	White
6	0.837430	1	White
7	0.773571	1	White
8	0.690866	1	Non-White
9	0.590255	1	White
10	0.468367	0	White
11	0.357743	1	Non-White
12	0.276027	0	White
13	0.193268	0	Non-White
14	0.145628	1	Non-White
15	0.103324	0	White
16	0.062255	0	Non-White
17	0.028513	0	White

Label: 1 positive, 0 negative

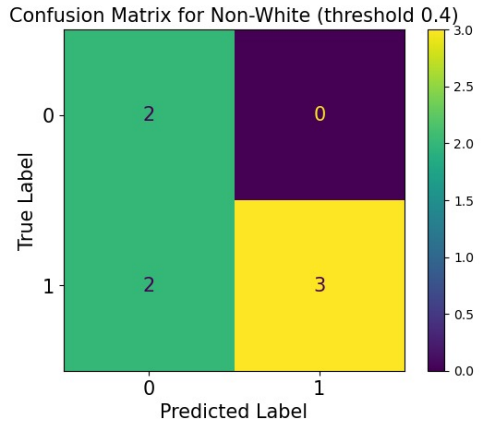
Fairness

Receiver operating curve

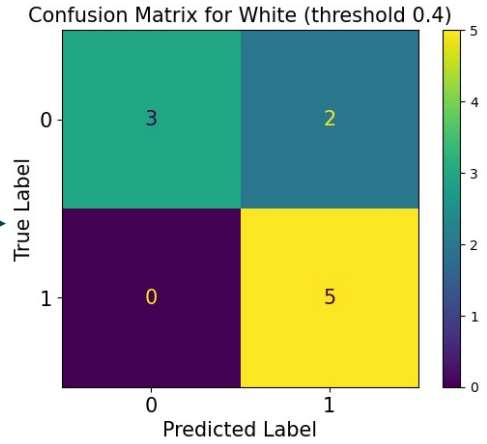
- Predict liver disease diagnostics

Index	Score	Label_value	Race
1	0.999800	1	White
2	0.986034	1	Non-White
3	0.965328	1	Non-White
4	0.927179	1	White
5	0.889940	0	White
6	0.837430	1	White
7	0.773571	1	White
8	0.690866	1	Non-White
9	0.590255	1	White
10	0.468367	0	White
11	0.357743	1	Non-White
12	0.276027	0	White
13	0.193268	0	Non-White
14	0.145628	1	Non-White
15	0.103324	0	White
16	0.062255	0	Non-White
17	0.028513	0	White

Label: 1 positive, 0 negative



TPR = tp / (tp + fn) = 3/(3+2) = 0.6
FPR = fp / (fp +tn) = 0/(0 + 2) = 0
Accuracy = 5/7 = 0.7



TPR = tp / (tp + fn) = 5/(5 + 0) = 1
FPR = fp / (fp +tn) = 2/(2+3) = 0.4
Accuracy = 8/10 = 0.8

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

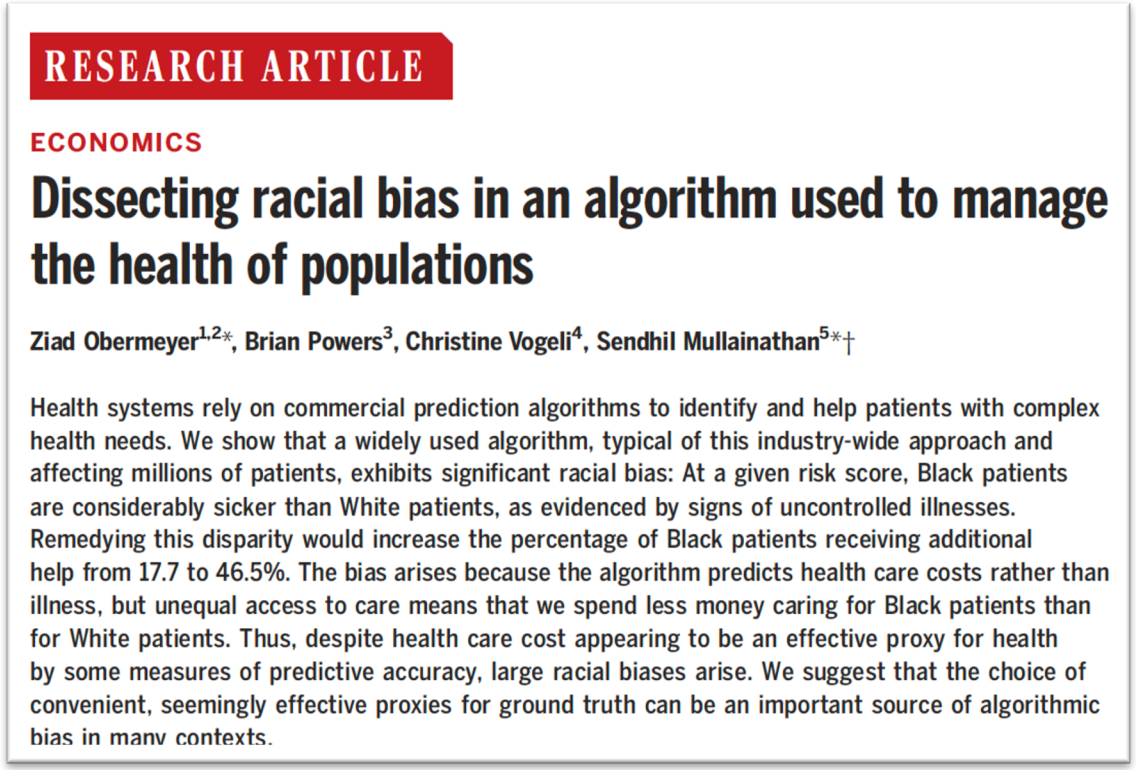
Fairness

Bias

- Bias is about disparate errors against specific sub-groups.

Use Case

Dissecting racial bias in an algorithm used to manage the health of populations



RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

Ziad Obermeyer *et al.* Dissecting racial bias in an algorithm used to manage the health of populations.*Science***366**,447-453(2019).
DOI:10.1126/science.aax2342

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Use Case

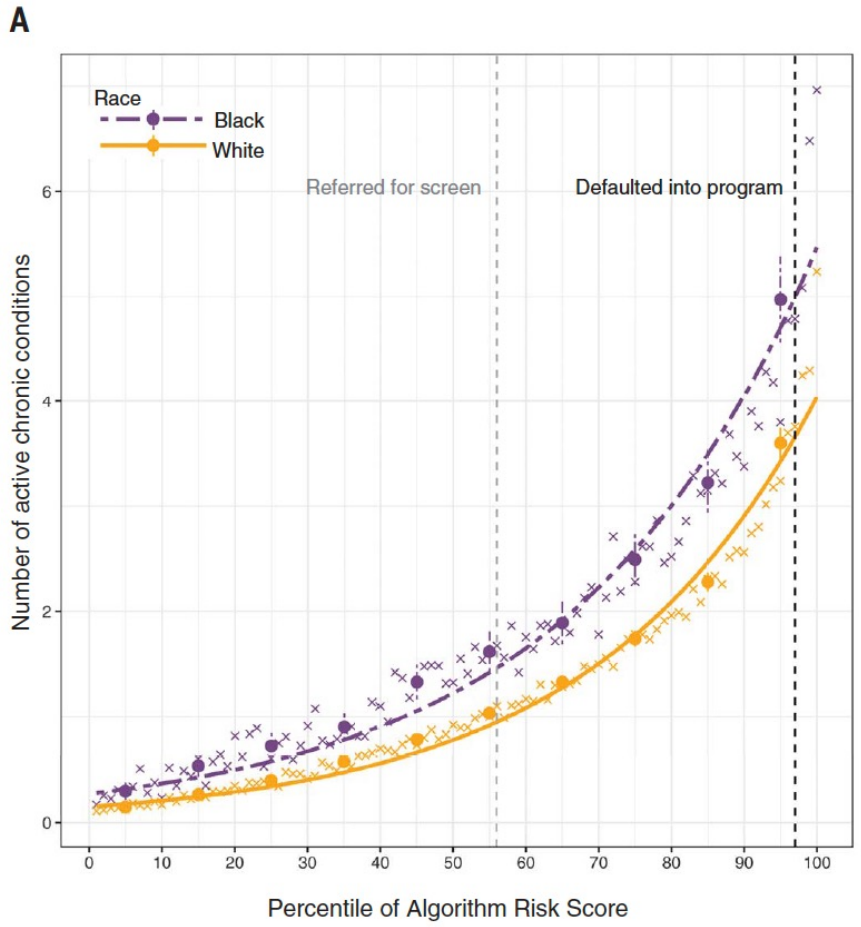
Dissecting racial bias in an algorithm used to manage the health of populations

- **Health System Data**: The study uses data from a commercial risk-prediction algorithm deployed in U.S. health systems to manage high-risk patients. The prediction is used to identify patients who should be enrolled in high-risk care management programs.
- **Study Sample**: It includes data from 43,539 White and 6,079 Black primary care patients enrolled in risk-based contracts from 2013 to 2015.
- **Algorithm Data**: The dataset contains the algorithm's predictions and the underlying data, including demographics, diagnoses, medications, and costs.
- **Outcome Measures**: The target variable in the article is future health care costs. Patients above the 97th percentile are automatically identified for enrollment in the program. Those above the 55th percentile are referred to their primary care physician and evaluates the impact of the program.
- **How does racial bias manifest in a widely used commercial prediction algorithm?**

Ziad Obermeyer *et al.* Dissecting racial bias in an algorithm used to manage the health of populations.*Science***366**,447-453(2019).
DOI:10.1126/science.aax2342

Use Case

Dissecting racial bias in an algorithm used to manage the health of populations

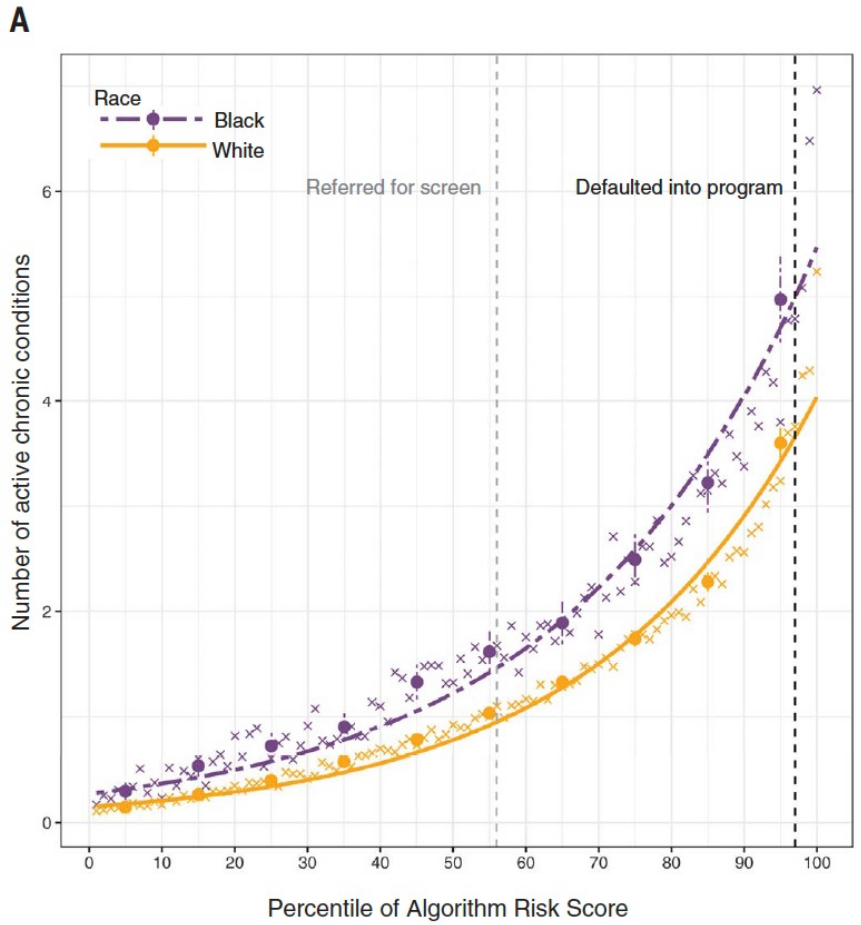


Ziad Obermeyer *et al.* Dissecting racial bias in an algorithm used to manage the health of populations.*Science***366**,447-453(2019).
DOI:10.1126/science.aax2342

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Use Case

Dissecting racial bias in an algorithm used to manage the health of populations



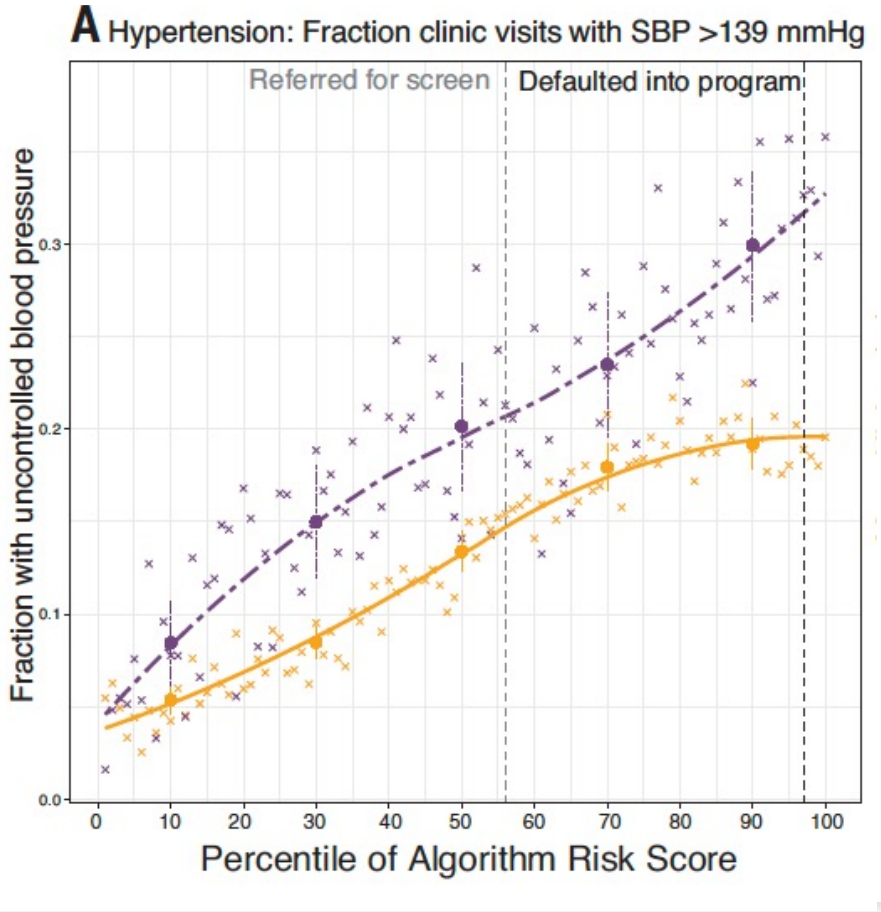
- At the same level of algorithm-predicted risk, Black patients have a significantly higher burden of chronic illnesses compared to White patients.
- This discrepancy indicates that the algorithm's risk scores do not accurately reflect the actual health status of Black patients, leading to underestimation of their health needs.

Ziad Obermeyer *et al.* Dissecting racial bias in an algorithm used to manage the health of populations.*Science***366**,447-453(2019).
DOI:10.1126/science.aax2342

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Use Case

Dissecting racial bias in an algorithm used to manage the health of populations

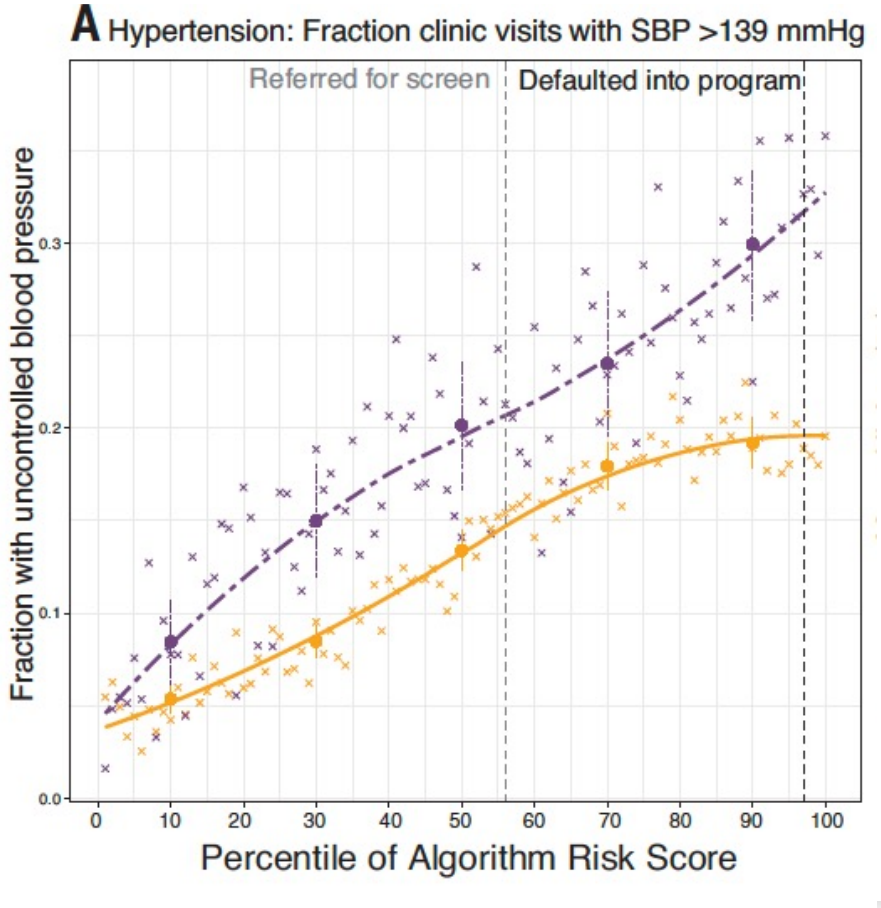


Ziad Obermeyer *et al.* Dissecting racial bias in an algorithm used to manage the health of populations.*Science***366**,447-453(2019).
DOI:10.1126/science.aax2342

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Use Case

Dissecting racial bias in an algorithm used to manage the health of populations



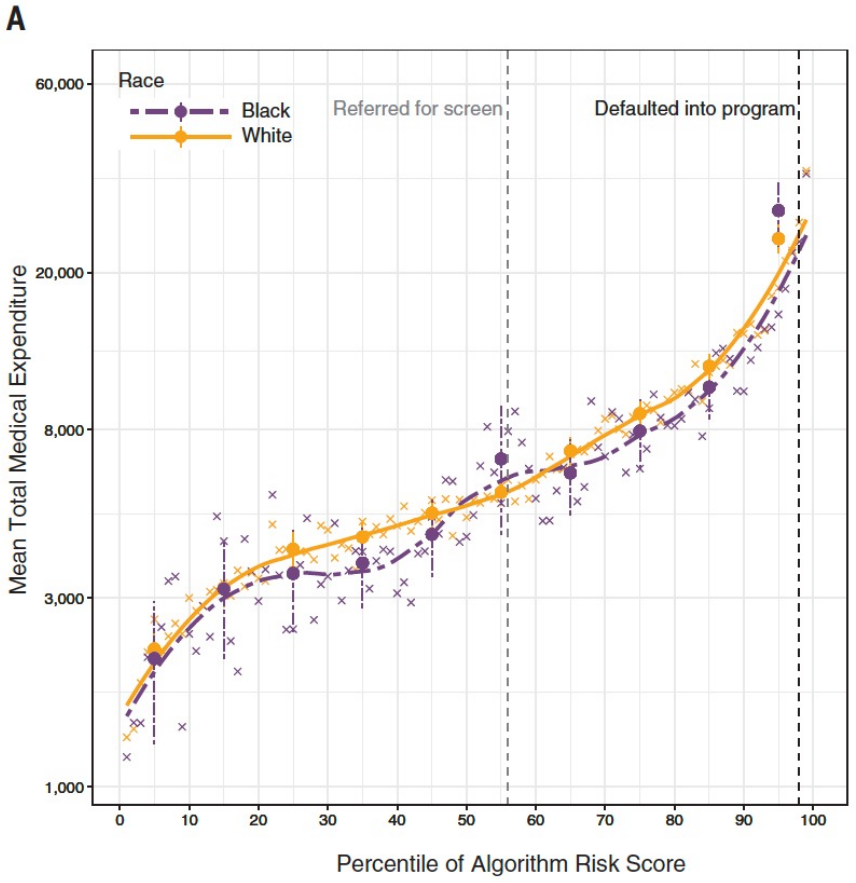
- This figure shows that Black patients have higher rates of uncontrolled hypertension (systolic blood pressure > 139 mmHg) than White patients at the same algorithm-predicted risk levels.
- This indicates that, despite similar risk scores, Black patients are experiencing more severe hypertension, suggesting that the algorithm underestimates the true health needs of Black patients.

Ziad Obermeyer *et al.* Dissecting racial bias in an algorithm used to manage the health of populations.*Science***366**,447-453(2019).
DOI:10.1126/science.aax2342

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Use Case

Dissecting racial bias in an algorithm used to manage the health of populations

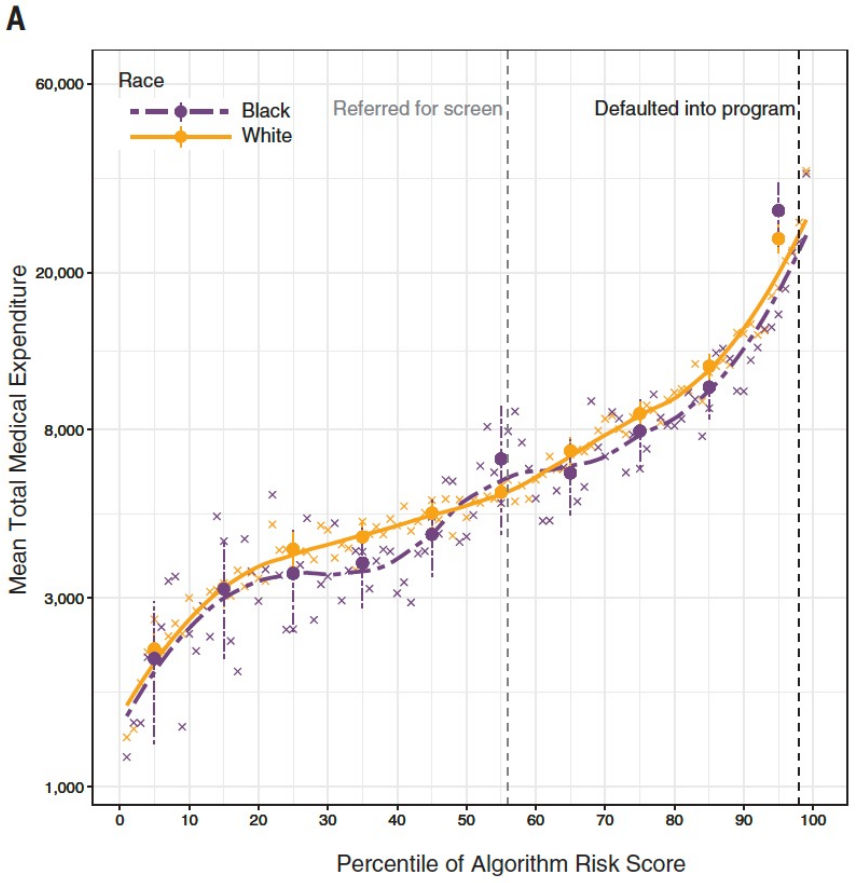


Ziad Obermeyer *et al.* Dissecting racial bias in an algorithm used to manage the health of populations.*Science***366**,447-453(2019).
DOI:10.1126/science.aax2342

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Use Case

Dissecting racial bias in an algorithm used to manage the health of populations



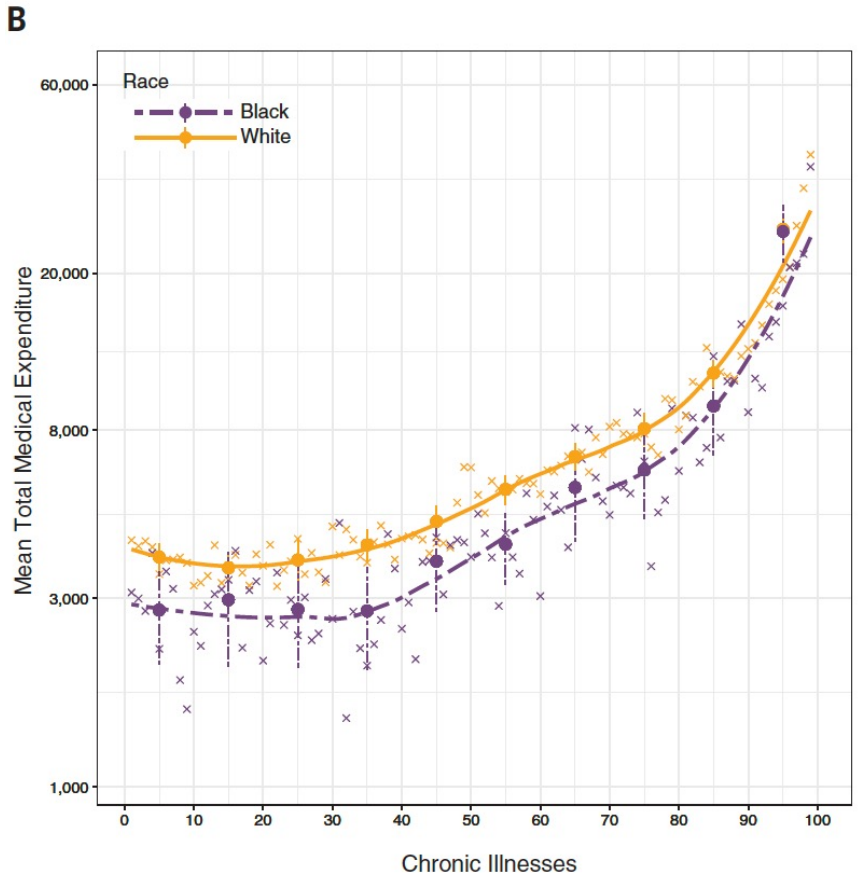
- Fig A shows that, at each level of predicted risk, Black and White patients have similar total medical expenditures.
- This suggests that, in terms of cost predictions, the algorithm appears well-calibrated across races.

Ziad Obermeyer *et al.* Dissecting racial bias in an algorithm used to manage the health of populations.*Science***366**,447-453(2019).
DOI:10.1126/science.aax2342

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Use Case

Dissecting racial bias in an algorithm used to manage the health of populations

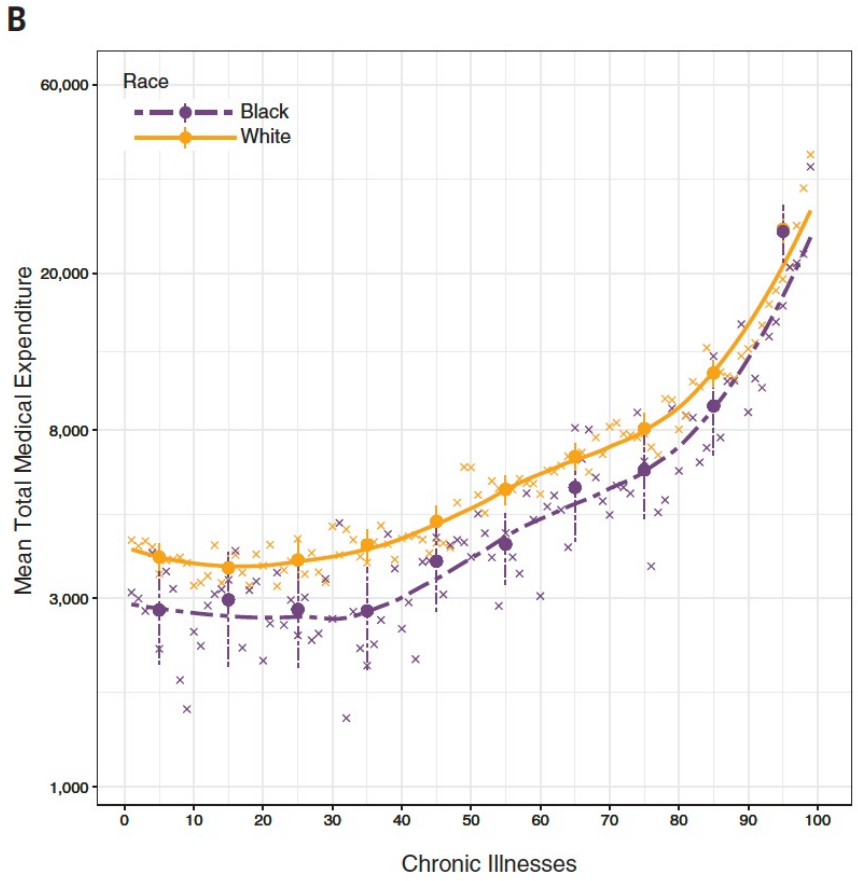


Ziad Obermeyer *et al.* Dissecting racial bias in an algorithm used to manage the health of populations.*Science***366**,447-453(2019).
DOI:10.1126/science.aax2342

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Use Case

Dissecting racial bias in an algorithm used to manage the health of populations



- The algorithm's reliance on cost as a proxy for health leads to racial bias.
- Because Black patients, despite having similar or greater health needs, generate lower medical expenses due to systemic barriers to accessing care.

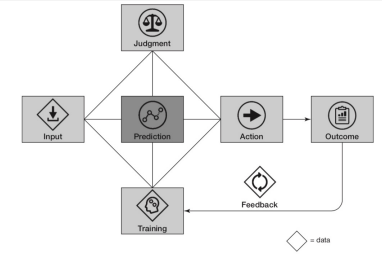
Where is bias coming from?

Ziad Obermeyer *et al.* Dissecting racial bias in an algorithm used to manage the health of populations.*Science***366**,447-453(2019).
DOI:10.1126/science.aax2342

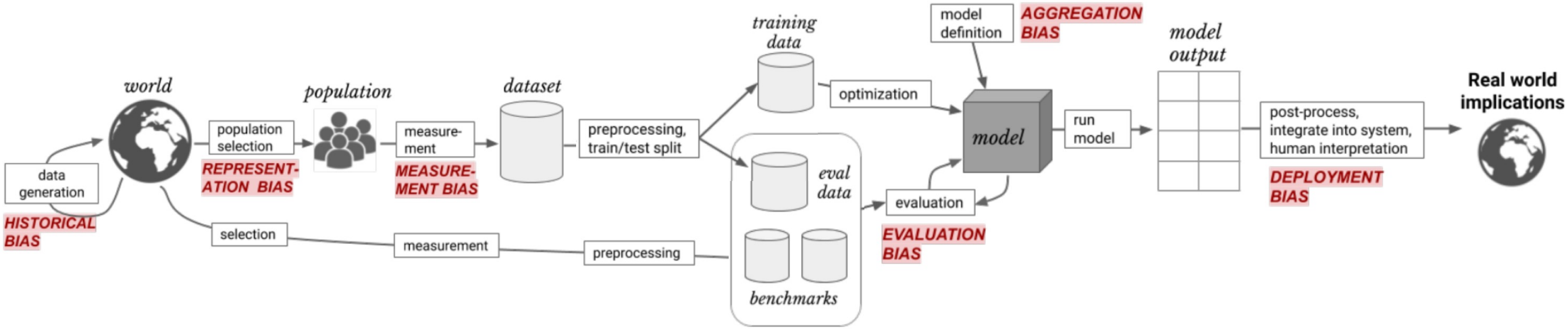
HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Fairness

Bias in decision making



- Bias is about disparate errors against specific sub-groups.

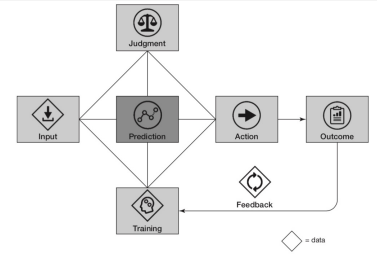


https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai

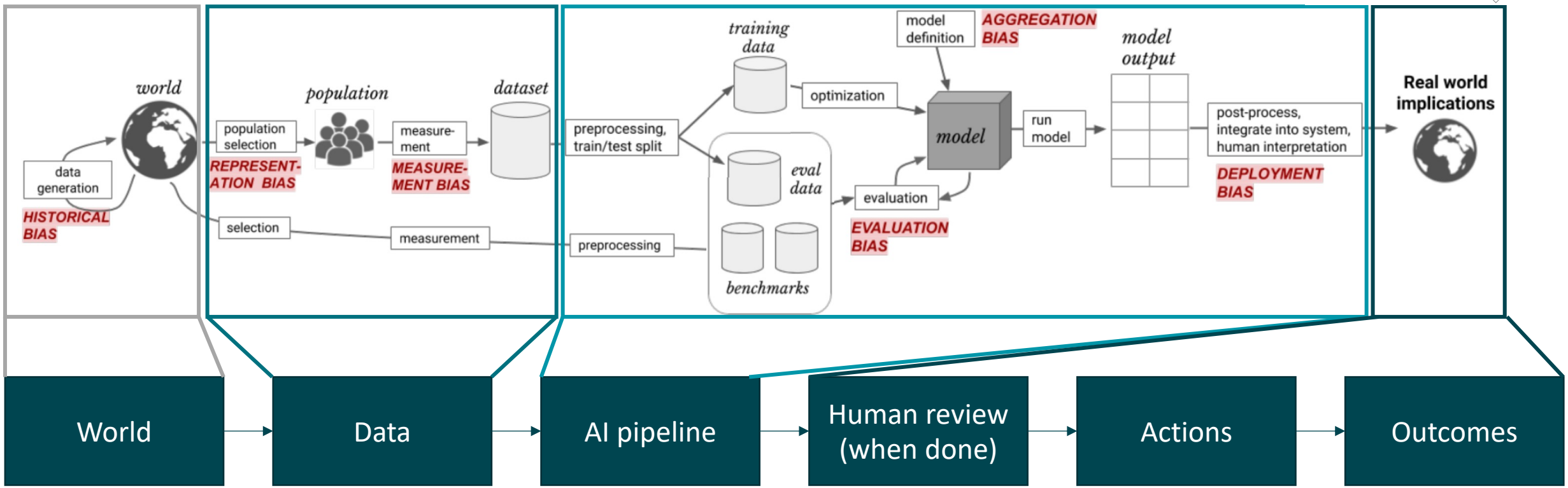
HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Fairness

Bias in decision making



- Bias is about disparate errors against specific sub-groups.



https://www.kaggle.com/code/alexisbcook/identifying-bias-in-ai

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Fairness

How to measure bias?

		Predicted condition			
	Total population = P + N	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) = TPR + TNR − 1	Prevalence threshold (PT) = $\frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power = $\frac{TP}{P}$ = 1 − FNR	False negative rate (FNR), miss rate = $\frac{FN}{P}$ = 1 − TPR
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection	False positive rate (FPR), probability of false alarm, fall-out = $\frac{FP}{N}$ = 1 − TNR	True negative rate (TNR), specificity (SPC), selectivity = $\frac{TN}{N}$ = 1 − FPR
	Prevalence = $\frac{P}{P+N}$	Positive predictive value (PPV), precision = $\frac{TP}{PP}$ = 1 − FDR	False omission rate (FOR) = $\frac{FN}{PN}$ = 1 − NPV	Positive likelihood ratio (LR+) = $\frac{TPR}{FPR}$	Negative likelihood ratio (LR−) = $\frac{FNR}{TNR}$
	Accuracy (ACC) = $\frac{TP+TN}{P+N}$	False discovery rate (FDR) = $\frac{FP}{PP}$ = 1 − PPV	Negative predictive value (NPV) = $\frac{TN}{PN}$ = 1 − FOR	Markedness (MK), deltaP (Δp) = PPV + NPV − 1	Diagnostic odds ratio (DOR) = $\frac{LR+}{LR-}$
	Balanced accuracy (BA) = $\frac{TPR+TNR}{2}$	F_1 score = $\frac{2PPV \times TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN}$	Fowlkes–Mallows index (FM) = $\sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) = $\sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index = $\frac{TP}{TP+FN+FP}$

Sources: [1][2][3][4][5][6][7][8][9] view · talk · edit

- Look for ration between groups

E.g.

$$Disparity_{FNR} = \frac{FNR_{black}}{FNR_{white}}$$

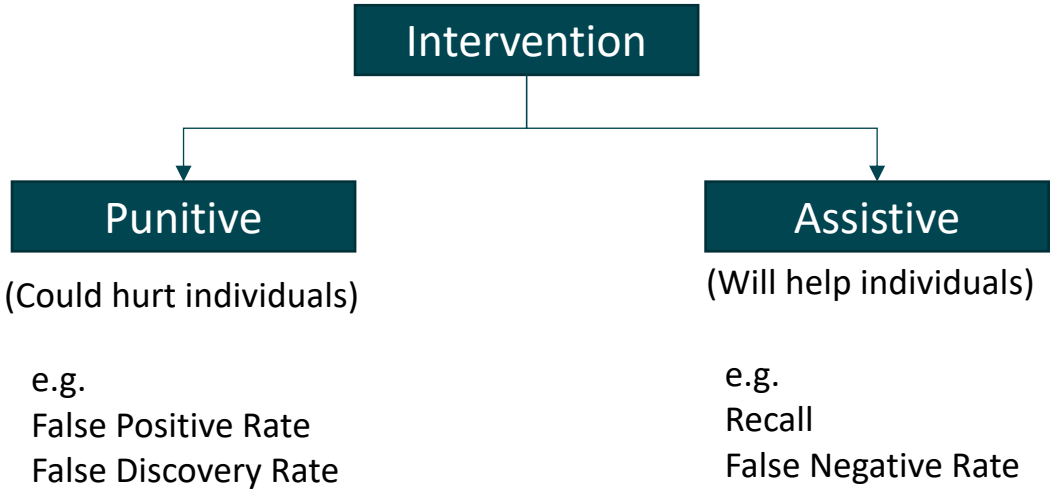
HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Fairness

Fairness depends on the intervention to be done

How to define the metric to use?

- If your intervention is punitive in nature, individuals may be harmed by intervening on them in error so you may care more about metrics that focus on false positives.
  - E.g. determining whom to deny a treatment
- If your intervention is assistive in nature, individuals may be harmed by failing to intervene on them when they have need, so you may care more about metrics that focus on false negatives.
  - E.g. determining who should receive an exercise voucher



Check the book for the full tree ☺ I am simplifying things here

https://textbook.coleridgeinitiative.org/chap-bias.html

Fairness

How to mitigate bias?

Pre-processing algorithms

- Relabeling
- Learned fair representation
- …

In-processing algorithms

- Regularization
- Constrained optimization
- …

Post-processing algorithms

- Threshold calibration
- …

https://textbook.coleridgeinitiative.org/chap-bias.html

Audit a model's predictions

Requirements

- Predictions or decisions
- Labels
- Sensitive attribute(s)

```
audit = Audit(df)
audit.summary_plot(metrics)
```

Correct a model's predictions

Requirements

- Model object
- Test set
- Validation set (recommended)

```
thresh = BalancedGroupThreshold(conf)
thresh.fit(X, preds_val, y, s)
thresh.transform(X, preds_test, s)
```

Training models w/ fairness considerations

Requirements

- Dataset
- Model configs (optional)

```
experiment = DefaultExperiment(df)
experiment.run()
```

http://aequitas.dssg.io/
https://github.com/dssg/aequitas

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

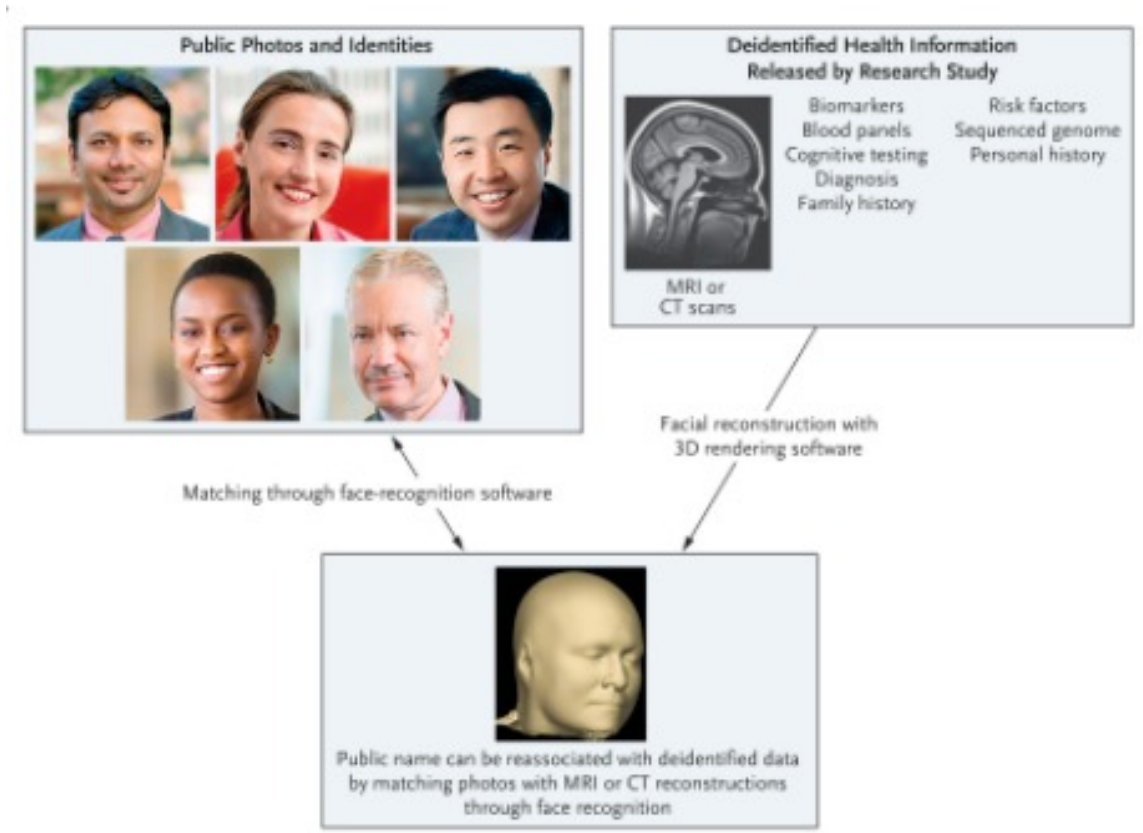
Privacy

Patient privacy should be a cornerstone principle while developing AI tools in medicine

https://gdpr.eu/

Privacy

Patient privacy should be a cornerstone principle while developing AI tools in medicine



Schwarz CG, et al. Identification of Anonymous MRI Research Participants with Face-Recognition Software. N Engl J Med. 2019 Oct 24;381(17):1684-1686. doi: 10.1056/NEJMc1908881.

Privacy

Patient privacy should be a cornerstone principle while developing AI tools in medicine

RESEARCH ARTICLE | NEUROSCIENCE

When makes you unique: Temporality of the human brain fingerprint

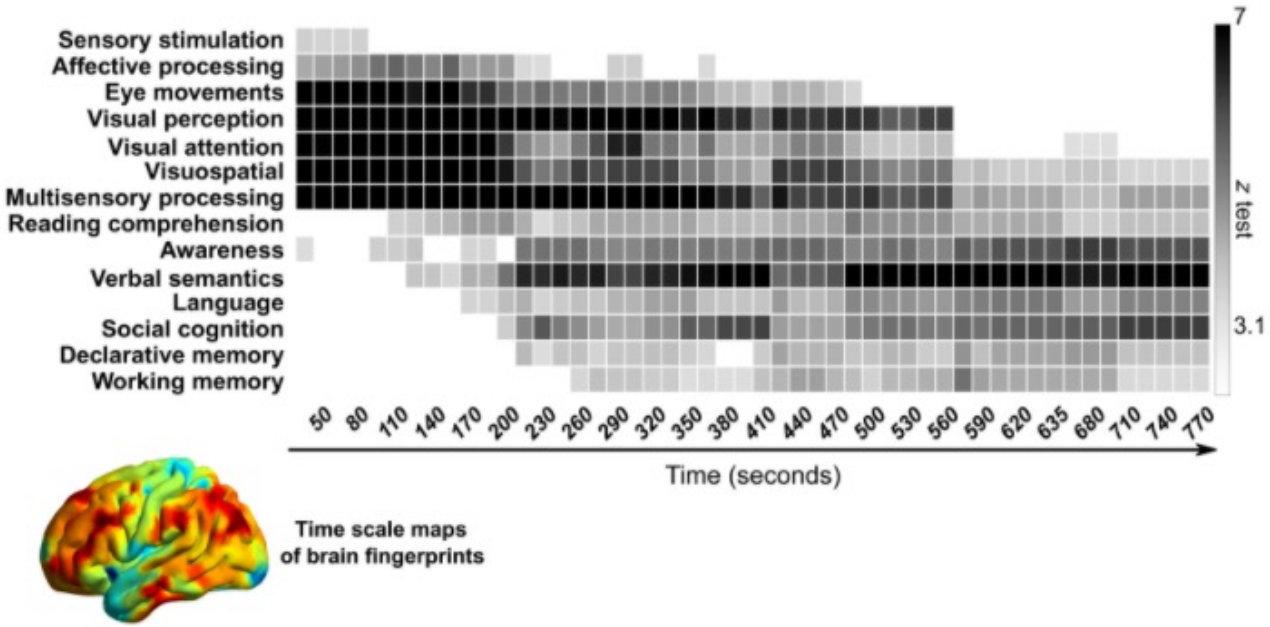
DIMITRI VAN DE VILLE, YOUNES FAROUJ, MARIA GIULIA PRETI, RAPHAËL LIÉGEOIS, AND ENRICO AMICO Authors Info & Affiliations

SCIENCE ADVANCES • 15 Oct 2021 • Vol 7, Issue 42 • DOI: 10.1126/sciadv.abj0751

14,219 36

Abstract

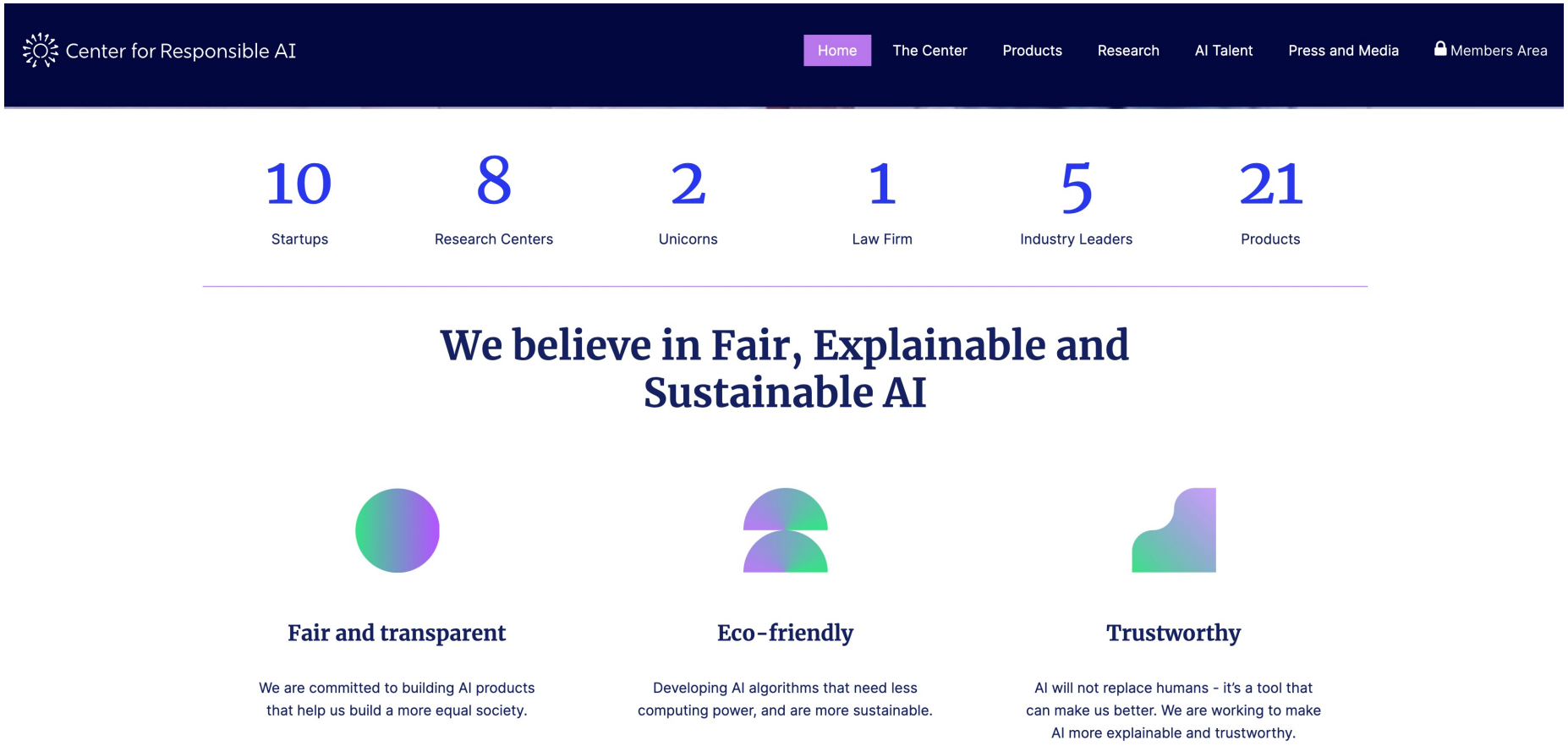
The extraction of "fingerprints" from human brain connectivity data has become a new frontier in neuroscience. However, the time scales of human brain identifiability are still largely unexplored. We here investigate the dynamics of brain fingerprints along two complementary axes: (i) What is the optimal time scale at which brain fingerprints integrate information and (ii) when best identification happens. Using dynamic identifiability, we show that the best identification emerges at longer time scales; however, short transient "bursts of identifiability," associated with neuronal activity, persist even when looking at shorter functional interactions. Furthermore, we report evidence that different parts of connectome finger-



Van De Ville D, et al. When makes you unique: Temporality of the human brain fingerprint. Sci Adv. 2021 Oct 15;7(42):eabj0751. doi: 10.1126/sciadv.abj0751.

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

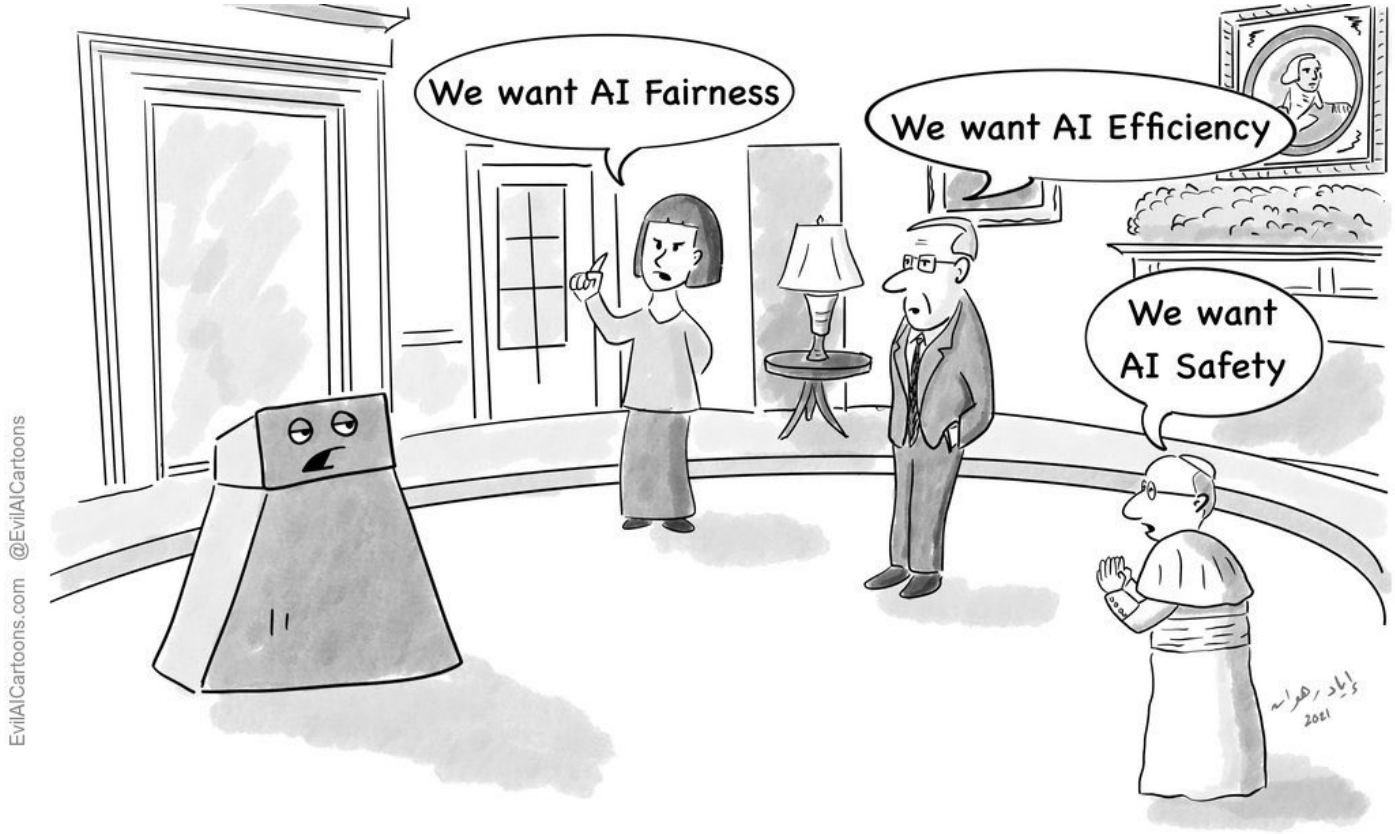
Are you interested in the topic? In Portugal we are part of Center for Responsible AI aiming to set standards to use AI responsibly



https://centerforresponsible.ai/

HOSPITAL DA LUZ LEARNING HEALTH
TRAINING, RESEARCH & INNOVATION CENTER

Questions?



<< And I want infinite battery! Talk to me when you've negotiated the tradeoffs! >>