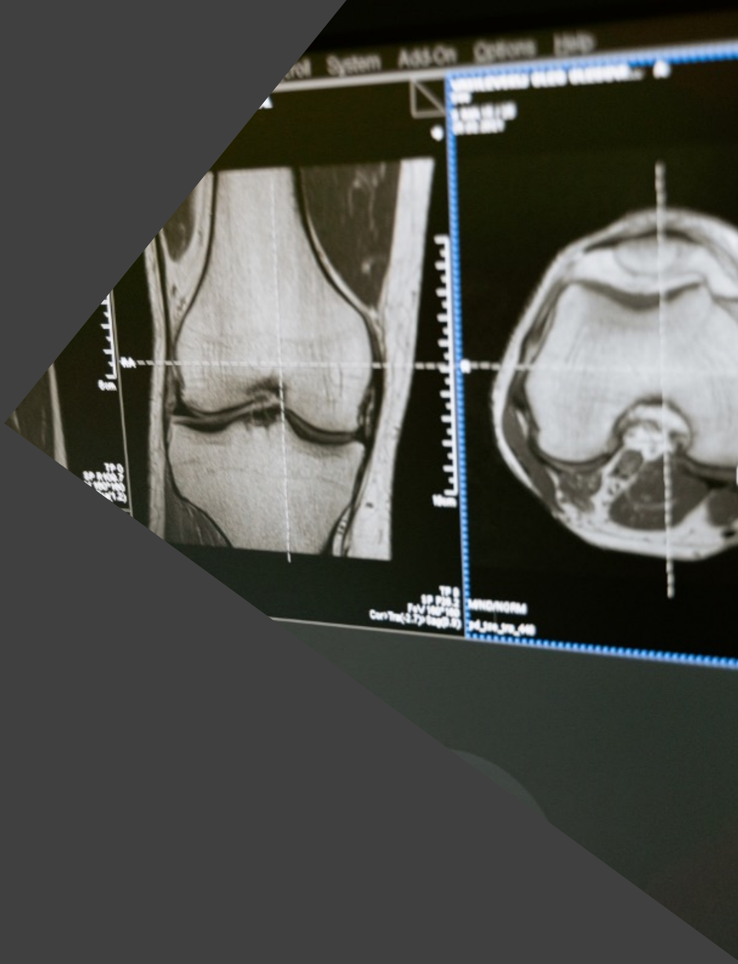
The Alan Turing Institute

The role of explainability to address equity challenges in AI-enabled medical devices



Antonella Perini

Research Associate

Ethics Team, Public Policy Programme

A rapid review on Equity in AI-enabled medical devices

A rapid review

What is the extent and nature of the existing bias in AI-enabled medical devices?

What are the existing guidelines, recommendations, and regulations to reduce or prevent the risk of such bias?

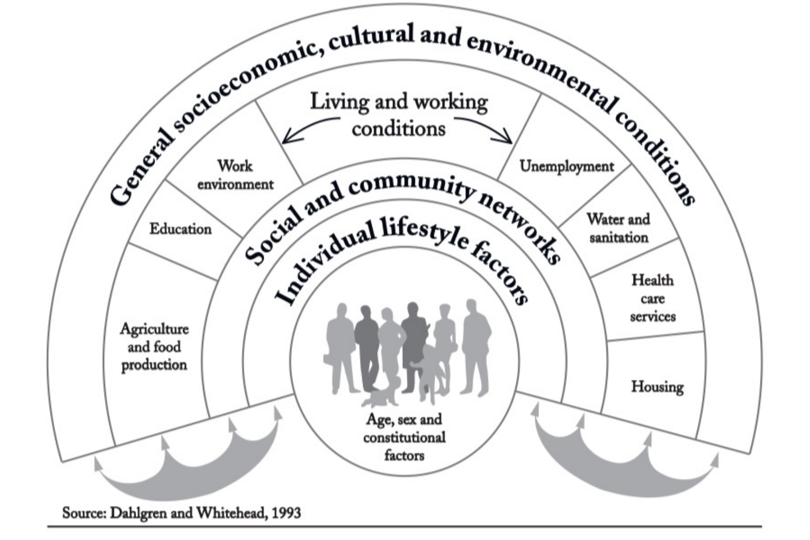
Key concepts

Health equity

A broad principle that signals a commitment to reduce or eliminate disparities in health by addressing the systemic challenges to achieving it

Social determinants of health

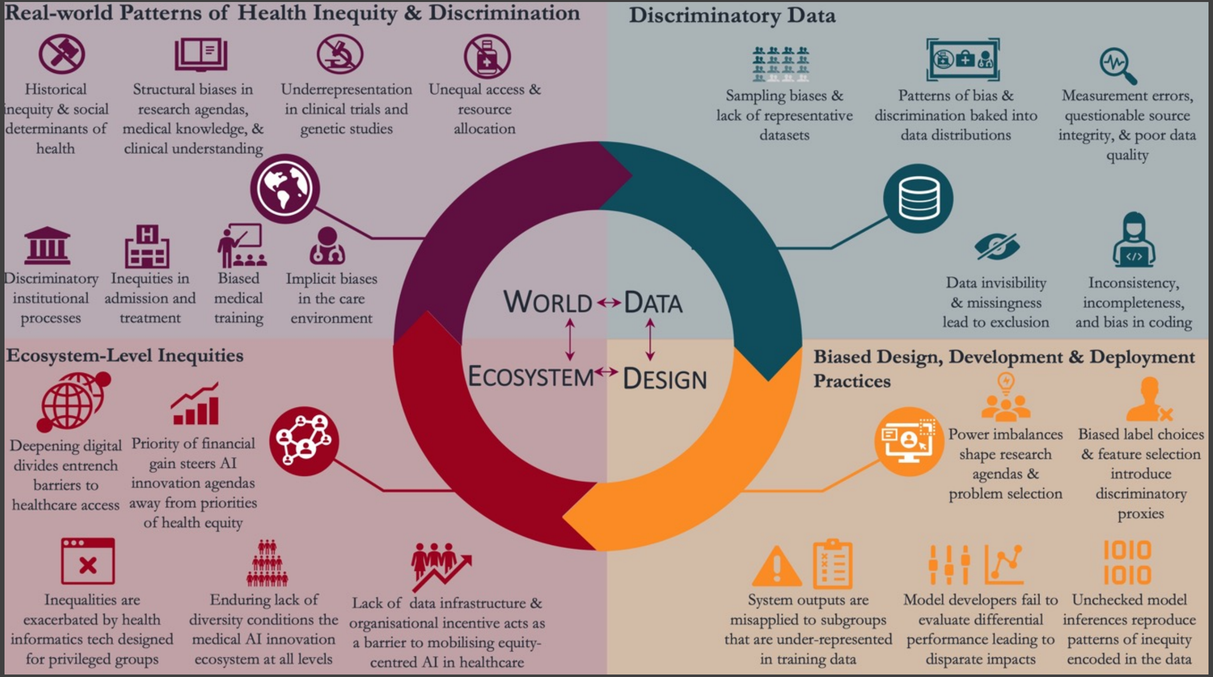
Non-medical factors that influence health outcomes


General socioeconomic, cultural and environmental conditions
Living and working conditions
Work environment
Education
Agriculture and food production
Unemployment
Water and sanitation
Health care services
Housing
Social and community networks
Individual lifestyle factors
Age, sex and constitutional factors

Source: Dahlgren and Whitehead, 1993

Dahlgren & Whitehead, 2006

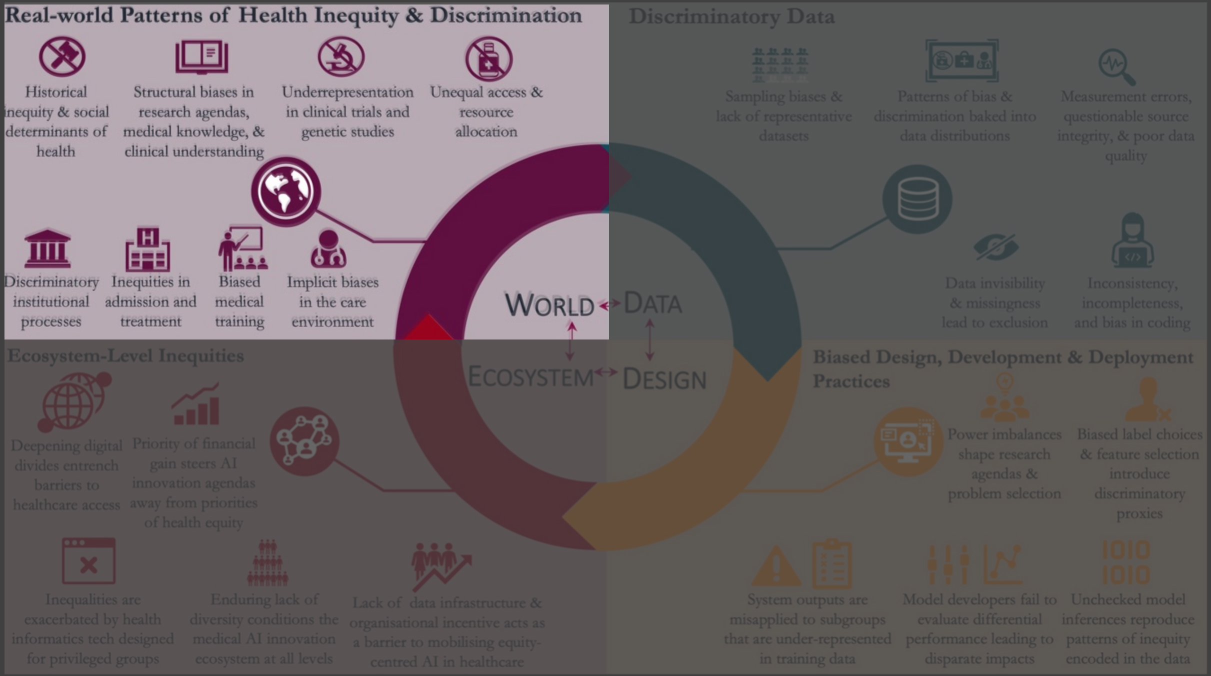
The Unvirtuous Circle


Real-world Patterns of Health Inequity & Discrimination
Historical inequity & social determinants of health
Structural biases in research agendas, medical knowledge, & clinical understanding
Underrepresentation in clinical trials and genetic studies
Unequal access & resource allocation
Discriminatory institutional processes
Inequities in admission and treatment
Biased medical training
Implicit biases in the care environment
Discriminatory Data
Sampling biases & lack of representative datasets
Patterns of bias & discrimination baked into data distributions
Measurement errors, questionable source integrity, & poor data quality
Data invisibility & missingness lead to exclusion
Inconsistency, incompleteness, and bias in coding
WORLD ↔ DATA
ECOSYSTEM ↔ DESIGN
Ecosystem-Level Inequities
Deepening digital divides entrench barriers to healthcare access
Priority of financial gain steers AI innovation agendas away from priorities of health equity
Inequalities are exacerbated by health informatics tech designed for privileged groups
Enduring lack of diversity conditions the medical AI innovation ecosystem at all levels
Lack of data infrastructure & organisational incentive acts as a barrier to mobilising equity-centred AI in healthcare
Biased Design, Development & Deployment Practices
Power imbalances shape research agendas & problem selection
Biased label choices & feature selection introduce discriminatory proxies
System outputs are misapplied to subgroups that are under-represented in training data
Model developers fail to evaluate differential performance leading to disparate impacts
Unchecked model inferences reproduce patterns of inequity encoded in the data
1010
1010

Cascading effects of health inequity across the AIEMD ecosystem (adapted from Leslie et al., 2021)

We employed a four-quadrant model, inspired by the one developed by Leslie et al. (2021) to conceptualise the key literature and illustrate how bias and inequity enter into the medical technology ecosystem as a cascading effect.

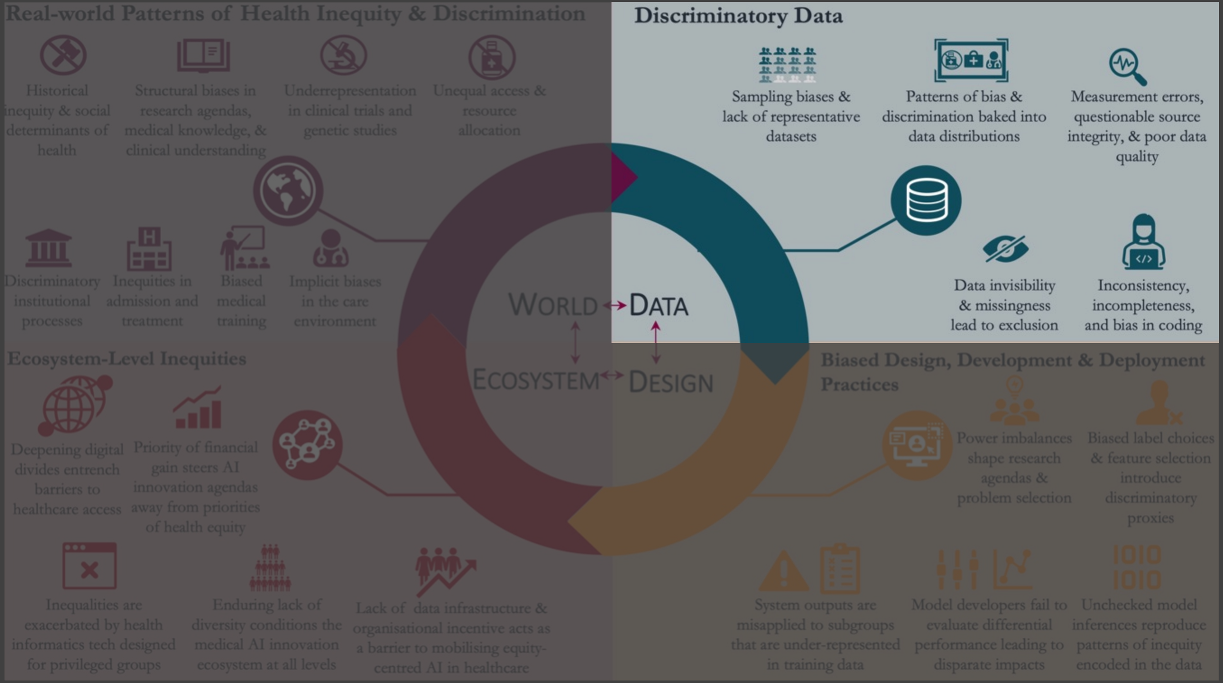
The Unvirtuous Circle


Real-world Patterns of Health Inequity & Discrimination
Historical inequity & social determinants of health
Structural biases in research agendas, medical knowledge, & clinical understanding
Underrepresentation in clinical trials and genetic studies
Unequal access & resource allocation
Discriminatory institutional processes
Inequities in admission and treatment
Biased medical training
Implicit biases in the care environment
Discriminatory Data
Sampling biases & lack of representative datasets
Patterns of bias & discrimination baked into data distributions
Measurement errors, questionable source integrity, & poor data quality
Data invisibility & missingness lead to exclusion
Inconsistency, incompleteness, and bias in coding
WORLD
DATA
ECOSYSTEM
DESIGN
Ecosystem-Level Inequities
Deepening digital divides entrench barriers to healthcare access
Priority of financial gain steers AI innovation agendas away from priorities of health equity
Inequalities are exacerbated by health informatics tech designed for privileged groups
Enduring lack of diversity conditions the medical AI innovation ecosystem at all levels
Lack of data infrastructure & organisational incentive acts as a barrier to mobilising equity-centred AI in healthcare
Biased Design, Development & Deployment Practices
Power imbalances shape research agendas & problem selection
Biased label choices & feature selection introduce discriminatory proxies
System outputs are misapplied to subgroups that are under-represented in training data
Model developers fail to evaluate differential performance leading to disparate impacts
Unchecked model inferences reproduce patterns of inequity encoded in the data

Cascading effects of health inequity across the AIEMD ecosystem (adapted from Leslie et al., 2021)

The **World** represents the social and historical reality that precedes the design of the medical devices. Failure to consider real-world patterns of inequality and discrimination in an AI project can lead to inequity at every stage of the AI lifecycle.

The Unvirtuous Circle



Real-world Patterns of Health Inequity & Discrimination

Historical inequity & social determinants of health

Structural biases in research agendas, medical knowledge, & clinical understanding

Underrepresentation in clinical trials and genetic studies

Unequal access & resource allocation

Discriminatory institutional processes

Inequities in admission and treatment

Biased medical training

Implicit biases in the care environment

Discriminatory Data

Sampling biases & lack of representative datasets

Patterns of bias & discrimination baked into data distributions

Measurement errors, questionable source integrity, & poor data quality

Data invisibility & missingness lead to exclusion

Inconsistency, incompleteness, and bias in coding

Ecosystem-Level Inequities

Deepening digital divides entrench barriers to healthcare access

Priority of financial gain steers AI innovation agendas away from priorities of health equity

Inequalities are exacerbated by health informatics tech designed for privileged groups

Enduring lack of diversity conditions the medical AI innovation ecosystem at all levels

Lack of data infrastructure & organisational incentive acts as a barrier to mobilising equity-centred AI in healthcare

Biased Design, Development & Deployment Practices

Power imbalances shape research agendas & problem selection

Biased label choices & feature selection introduce discriminatory proxies

System outputs are misapplied to subgroups that are under-represented in training data

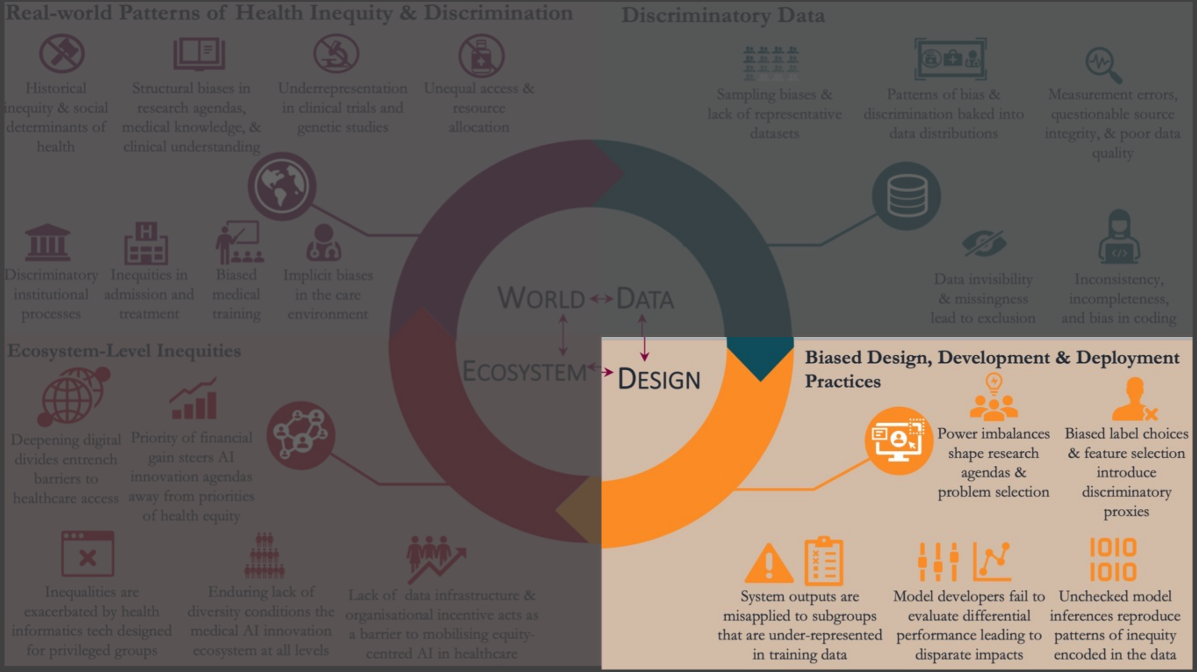
Model developers fail to evaluate differential performance leading to disparate impacts

Unchecked model inferences reproduce patterns of inequity encoded in the data

Cascading effects of health inequity across the AIEMD ecosystem (adapted from Leslie et al., 2021)

The **Data** represents the data used to train, test and/or validate the AI-supported medical technologies. Inequities can enter the data through various means, from insufficient representation of certain population groups, lack of data quality and completeness, to inadequate coding practices.

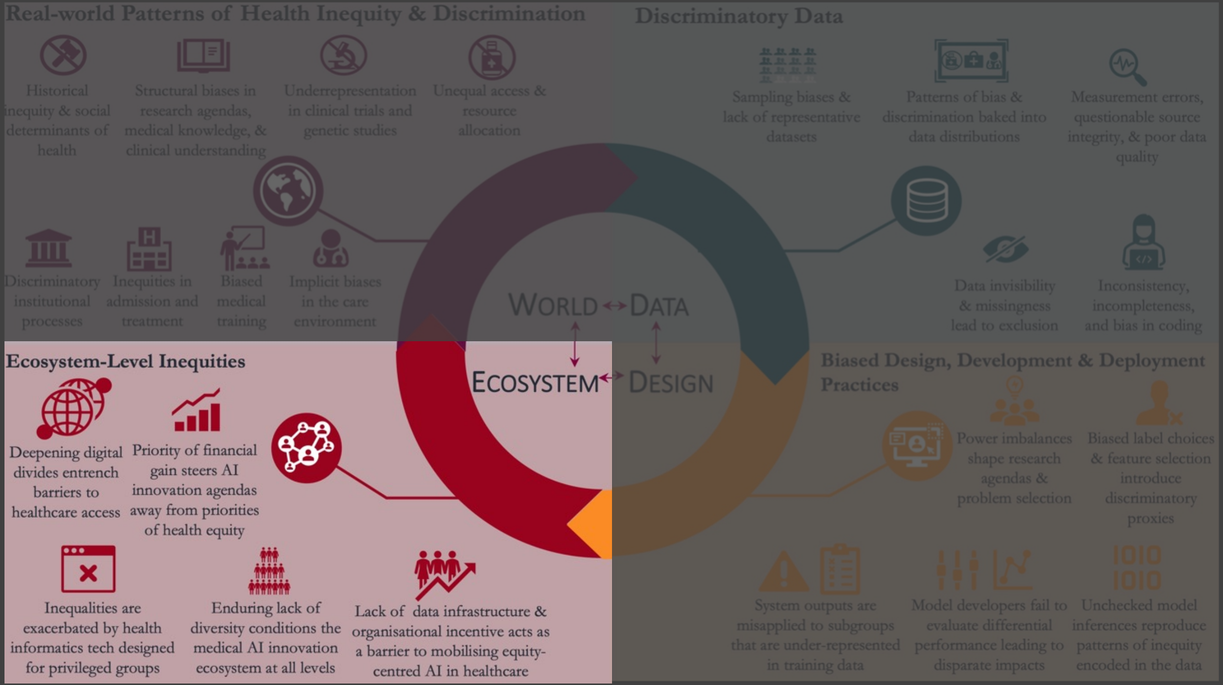
The Unvirtuous Circle


Real-world Patterns of Health Inequity & Discrimination
Historical inequity & social determinants of health
Structural biases in research agendas, medical knowledge, & clinical understanding
Underrepresentation in clinical trials and genetic studies
Unequal access & resource allocation
Discriminatory institutional processes
Inequities in admission and treatment
Biased medical training
Implicit biases in the care environment
Discriminatory Data
Sampling biases & lack of representative datasets
Patterns of bias & discrimination baked into data distributions
Measurement errors, questionable source integrity, & poor data quality
Data invisibility & missingness lead to exclusion
Inconsistency, incompleteness, and bias in coding
WORLD
DATA
ECOSYSTEM
DESIGN
Ecosystem-Level Inequities
Deepening digital divides entrench barriers to healthcare access
Priority of financial gain steers AI innovation agendas away from priorities of health equity
Inequalities are exacerbated by health informatics tech designed for privileged groups
Enduring lack of diversity conditions the medical AI innovation ecosystem at all levels
Lack of data infrastructure & organisational incentive acts as a barrier to mobilising equity-centred AI in healthcare
Biased Design, Development & Deployment Practices
Power imbalances shape research agendas & problem selection
Biased label choices & feature selection introduce discriminatory proxies
System outputs are misapplied to subgroups that are under-represented in training data
Model developers fail to evaluate differential performance leading to disparate impacts
Unchecked model inferences reproduce patterns of inequity encoded in the data

Cascading effects of health inequity across the AIEMD ecosystem (adapted from Leslie et al., 2021)

The **Design** represents the sociotechnical design, development, and deployment lifecycle of AI systems.

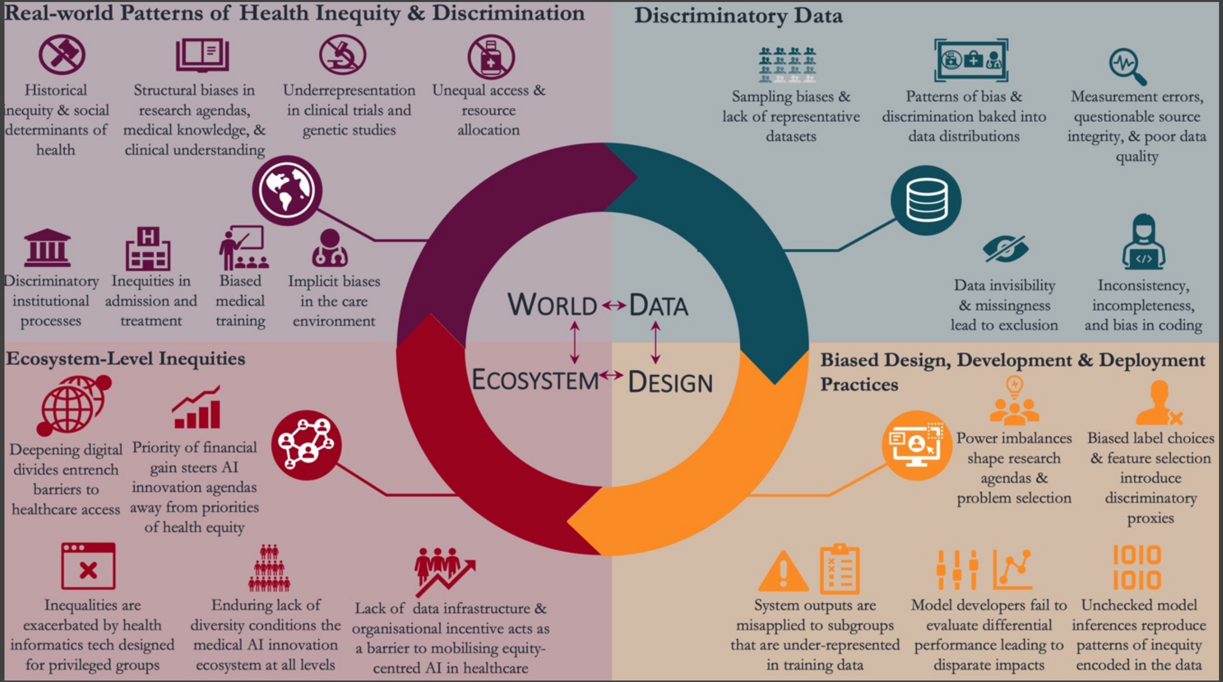
The Unvirtuous Circle


Real-world Patterns of Health Inequity & Discrimination
Historical inequity & social determinants of health
Structural biases in research agendas, medical knowledge, & clinical understanding
Underrepresentation in clinical trials and genetic studies
Unequal access & resource allocation
Discriminatory institutional processes
Inequities in admission and treatment
Biased medical training
Implicit biases in the care environment
Discriminatory Data
Sampling biases & lack of representative datasets
Patterns of bias & discrimination baked into data distributions
Measurement errors, questionable source integrity, & poor data quality
Data invisibility & missingness lead to exclusion
Inconsistency, incompleteness, and bias in coding
WORLD
DATA
ECOSYSTEM
DESIGN
Ecosystem-Level Inequities
Deepening digital divides entrench barriers to healthcare access
Priority of financial gain steers AI innovation agendas away from priorities of health equity
Inequalities are exacerbated by health informatics tech designed for privileged groups
Enduring lack of diversity conditions the medical AI innovation ecosystem at all levels
Lack of data infrastructure & organisational incentive acts as a barrier to mobilising equity-centred AI in healthcare
Biased Design, Development & Deployment Practices
Power imbalances shape research agendas & problem selection
Biased label choices & feature selection introduce discriminatory proxies
System outputs are misapplied to subgroups that are under-represented in training data
Model developers fail to evaluate differential performance leading to disparate impacts
Unchecked model inferences reproduce patterns of inequity encoded in the data

Cascading effects of health inequity across the AIEMD ecosystem (adapted from Leslie et al., 2021)

The **Ecosystem** represents the economic, legal, cultural, and political sphere. Inequities and biases at the Ecosystem level can steer or shape AI based medical innovation.

The Unvirtuous Circle



Real-world Patterns of Health Inequity & Discrimination

Historical inequity & social determinants of health

Structural biases in research agendas, medical knowledge, & clinical understanding

Underrepresentation in clinical trials and genetic studies

Unequal access & resource allocation

Discriminatory institutional processes

Inequities in admission and treatment

Biased medical training

Implicit biases in the care environment

Discriminatory Data

Sampling biases & lack of representative datasets

Patterns of bias & discrimination baked into data distributions

Measurement errors, questionable source integrity, & poor data quality

Data invisibility & missingness lead to exclusion

Inconsistency, incompleteness, and bias in coding

WORLD ↔ DATA

ECOSYSTEM ↔ DESIGN

Ecosystem-Level Inequities

Deepening digital divides entrench barriers to healthcare access

Priority of financial gain steers AI innovation agendas away from priorities of health equity

Inequalities are exacerbated by health informatics tech designed for privileged groups

Enduring lack of diversity conditions the medical AI innovation ecosystem at all levels

Lack of data infrastructure & organisational incentive acts as a barrier to mobilising equity-centred AI in healthcare

Biased Design, Development & Deployment Practices

Power imbalances shape research agendas & problem selection

Biased label choices & feature selection introduce discriminatory proxies

System outputs are misapplied to subgroups that are under-represented in training data

Model developers fail to evaluate differential performance leading to disparate impacts

Unchecked model inferences reproduce patterns of inequity encoded in the data

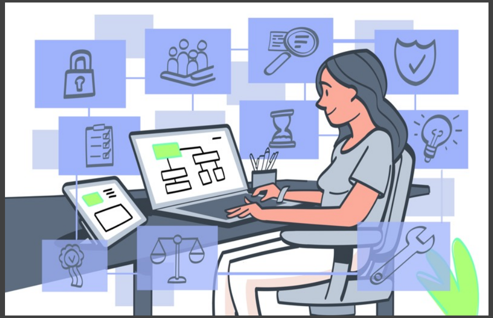
Cascading effects of health inequity across the AIEMD ecosystem (adapted from Leslie et al., 2021)

The model provides an upstream examination of real-world inequalities and the social determinants of health, which then flows into the construction of data that underlies AI-supported medical technologies, the sociotechnical features of system design, and the shaping influence and political economy of the medical technology ecosystem.

Explainability to address equity challenges

Explanations are critical to respond to equity challenges

Bias detection and mitigation



Compliance with regulation and standards

Risk management and redress

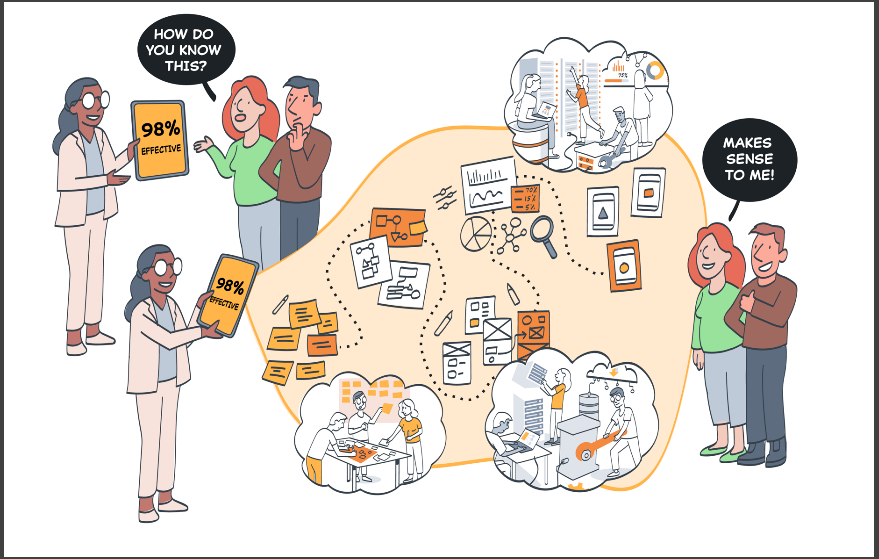
Meaningful and situated information for patients

Images by Jonny Lighthands, used under CC BY 4.0

The core principle of explainability

Explainability

the degree to which a system or set of tools support a person’s ability to explain and communicate the **behaviour of the system** or the **processes of designing, developing, and deploying the system** within a particular **context**.

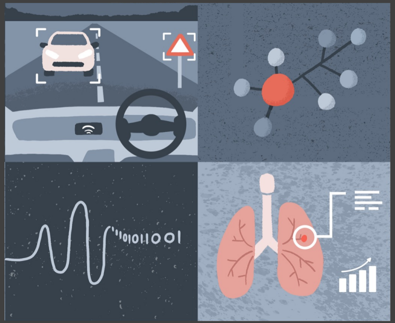


“Public trust in Science” by Jonny Lighthands, used under CC BY 4.0

Source: Burr, C., Fischer, C., and Rincon, C. (2023) Responsible Research and Innovation (Turing Commons Skills Track). Alan Turing Institute. 10.5281/zenodo.7755693.

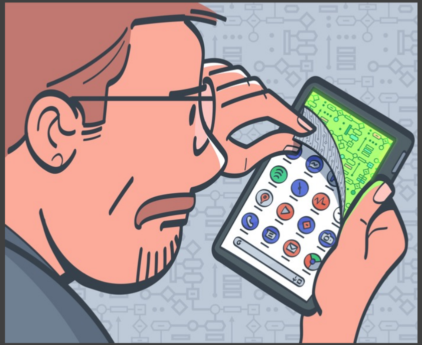
Three factors that support explainability

– Interpretable models



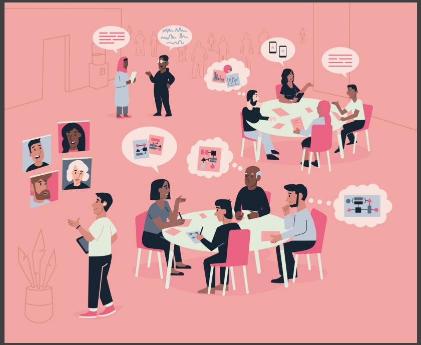
“Artificial Intelligence (Version 2)” by Jonny Lighthands, used under CC BY 4.0 / Cropped from original

– Project transparency



Transparency (with background)” by Jonny Lighthands, used under CC BY 4.0

– Situated explanations

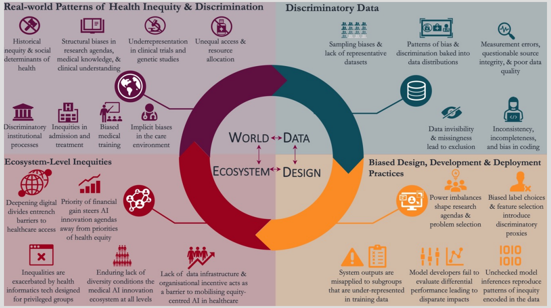


“Stakeholder Engageemnt” by Jonny Lighthands, used under CC BY 4.0

Source: Burr, C., Fischer, C., and Rincon, C. (2023) Responsible Research and Innovation (Turing Commons Skills Track). Alan Turing Institute. 10.5281/zenodo.7755693.

Wrap up

- Technical bias mitigation methods, but also ‘whole-of-society’ and multi-sectoral approaches to technology governance.
- Responsible research and innovation require a sociotechnical approach.
- The cascading effects of inequity will mean that explanations will also have compounding effects. This encourages an explanation-aware approach throughout the AI lifecycle.



Real-world Patterns of Health Inequity & Discrimination

Historical inequity & social determinants of health

Structural biases in research agendas, medical knowledge, & clinical understanding

Underrepresentation in clinical trials and genetic studies

Unequal access & resource allocation

Discriminatory institutional processes

Inequities in admission and treatment

Biased medical training

Implicit biases in the care environment

Discriminatory Data

Sampling biases & lack of representative datasets

Patterns of bias & discrimination baked into data distributions

Measurement errors, questionable source integrity, & poor data quality

Data invisibility & missingness lead to exclusion

Inconsistency, incompleteness, and bias in coding

WORLD ↔ DATA

ECOSYSTEM ↔ DESIGN

Ecosystem-Level Inequities

Deepening digital divides entrench barriers to healthcare access

Priority of financial gain steers AI innovation agendas away from priorities of health equity

Inequalities are exacerbated by health informatics tech designed for privileged groups

Enduring lack of diversity conditions the medical AI innovation ecosystem at all levels

Lack of data infrastructure & organisational incentive acts as a barrier to mobilising equity-centred AI in healthcare

Biased Design, Development & Deployment Practices

Power imbalances shape research agendas & problem selection

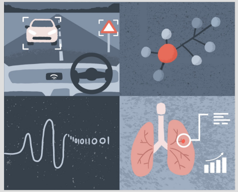
Biased label choices & feature selection introduce discriminatory proxies

System outputs are misapplied to subgroups that are under-represented in training data

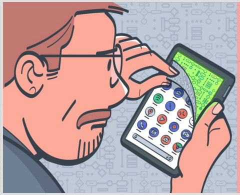
Model developers fail to evaluate differential performance leading to disparate impacts

Unchecked model inferences reproduce patterns of inequity encoded in the data

Interpretable models



Project transparency



Situated explanations

Thanks!

I would like to acknowledge our team on Equity in Medical Devices and Turing Commons for all of their input and work that went into this resource.

Please feel free to contact me with any additional questions.

Antonella Perini, aperini@turing.ac.uk

Useful resources

- **Responsible Research and Innovation in Data Science and AI (Turing Commons Skill Tracks)**: https://alan-turing-institute.github.io/turing-commons/skills-tracks/rri/index.html (Undergoing substantial updates)
- **Project ExplAIn:** https://www.turing.ac.uk/news/project-explain
- **Explaining decisions made with AI**: https://ico.org.uk/for-organisations/guide-to-data-protection/key-dp-themes/explaining-decisions-made-with-artificial-intelligence/
- **Public Policy Programme:** https://twitter.com/turingpubpol

The Alan Turing Institute