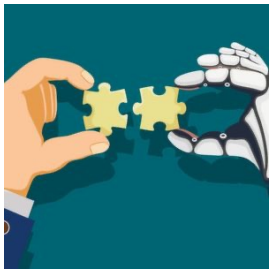


Improving Decision Making with Artificial Intelligence, Fairly

Manuel Gomez Rodriguez

Includes joint work with Nastaran Okati & Stratis Tsirtsis

MAX PLANCK INSTITUTE
FOR SOFTWARE SYSTEMS

MAX-PLANCK-GESELLSCHAFT

Artificial intelligence to Improve Decision Making

AI promises a new **generation** of **decision support systems** in many **high-stakes domains**

Health

Science

Education

Artificial Intelligence to Improve Decision Making

AI promises a new **generation** of **decision support systems** in many **high-stakes domains**

Health

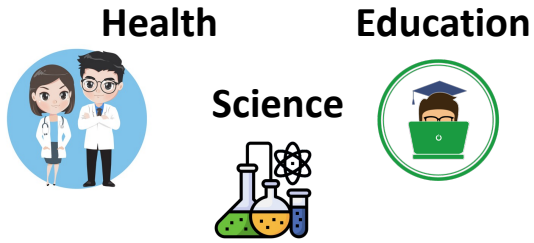
Education

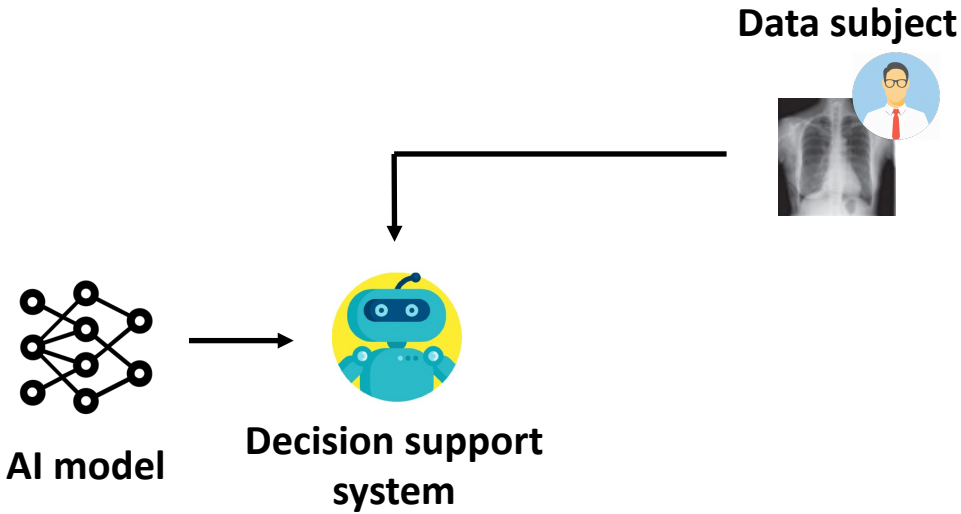
Science

Data subject

Artificial Intelligence to Improve Decision Making

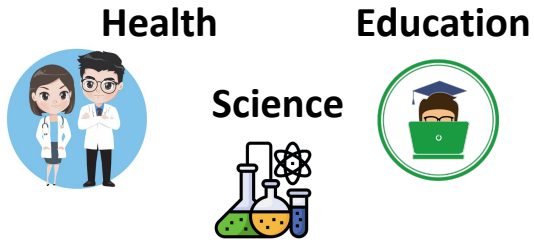
AI promises a new **generation** of **decision support systems** in many **high-stakes domains**

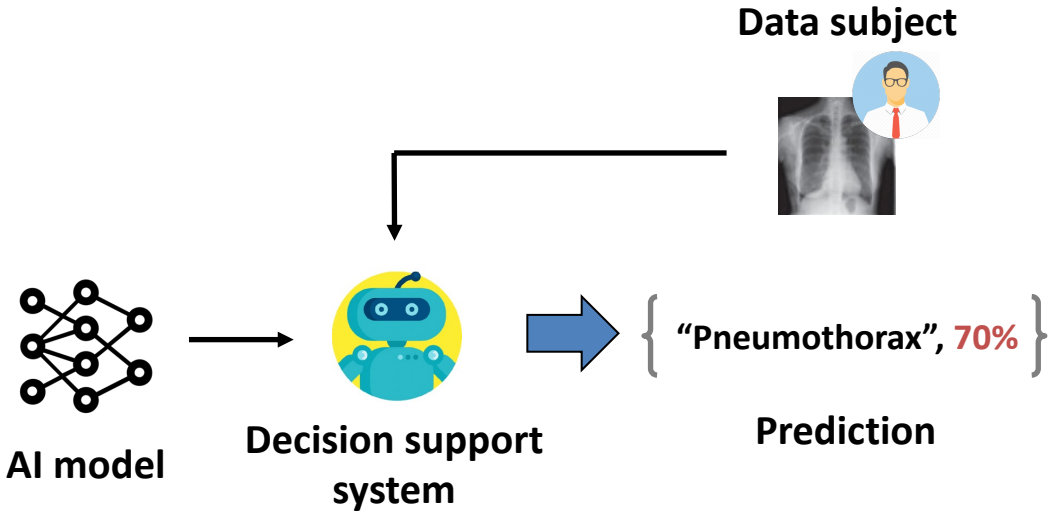




Artificial Intelligence to Improve Decision Making

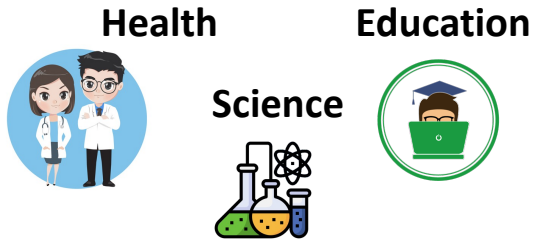
AI promises a new **generation** of **decision support systems** in many **high-stakes domains**

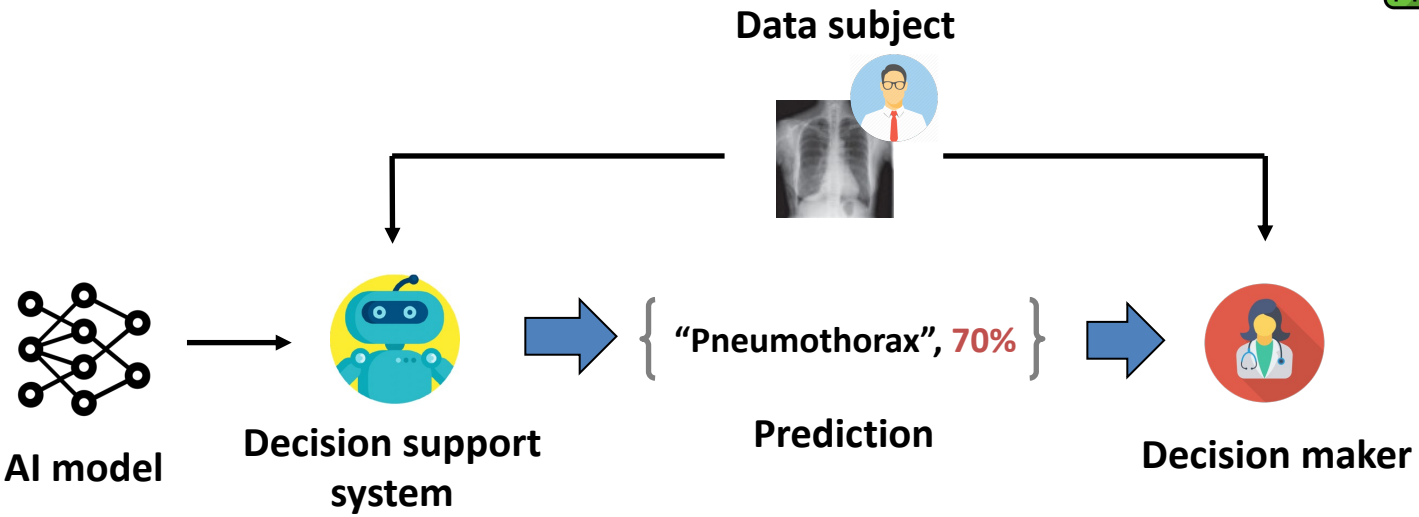




Artificial Intelligence to Improve Decision Making

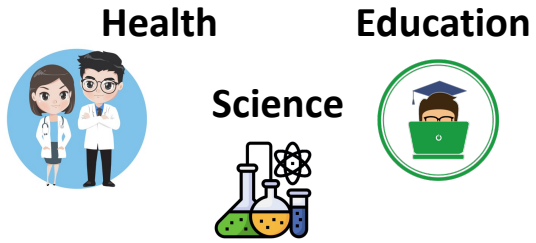
AI promises a new **generation** of **decision support systems** in many **high-stakes domains**

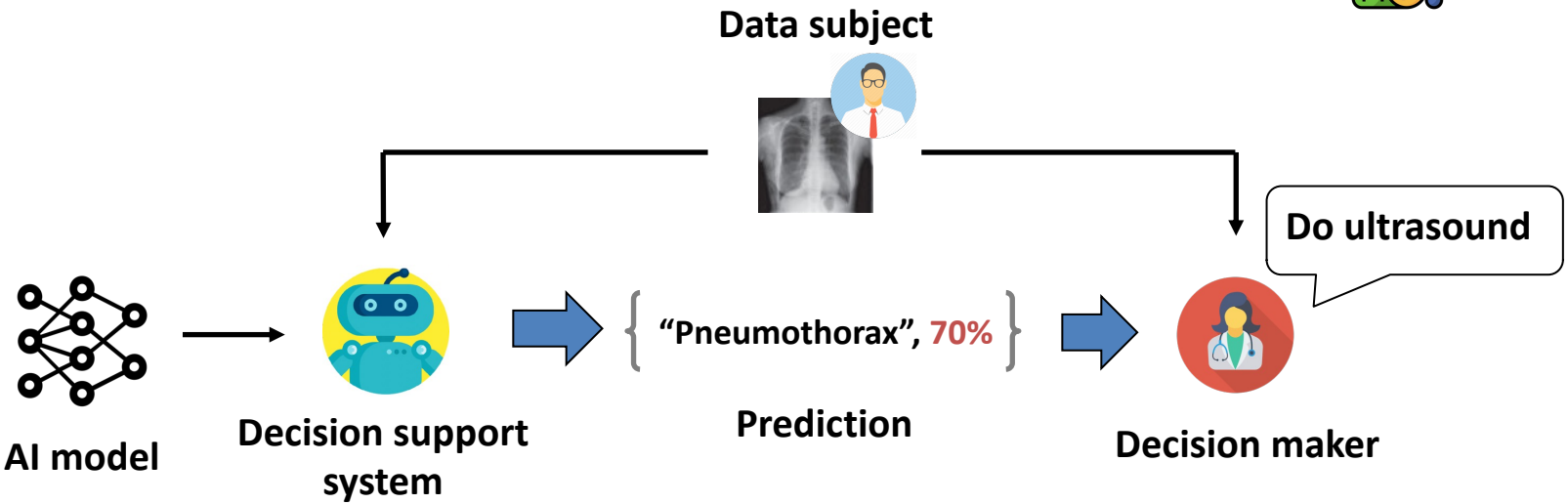




Artificial Intelligence to Improve Decision Making

AI promises a new **generation** of **decision support systems** in many **high-stakes domains**





Artificial Intelligence to Improve Decision Making

AI promises a new **generation** of **decision support systems** in many **high-stakes domains**

Health

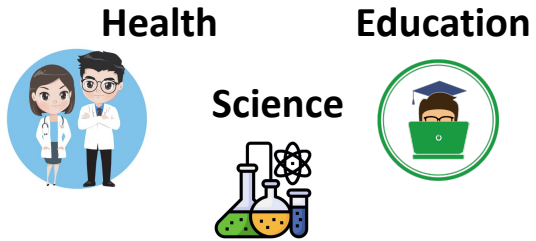
Science

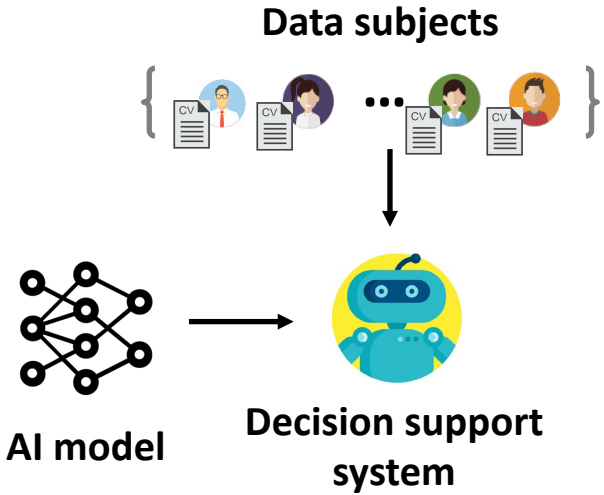
Education

Data subjects

Artificial Intelligence to Improve Decision Making

AI promises a new **generation** of **decision support systems** in many **high-stakes domains**





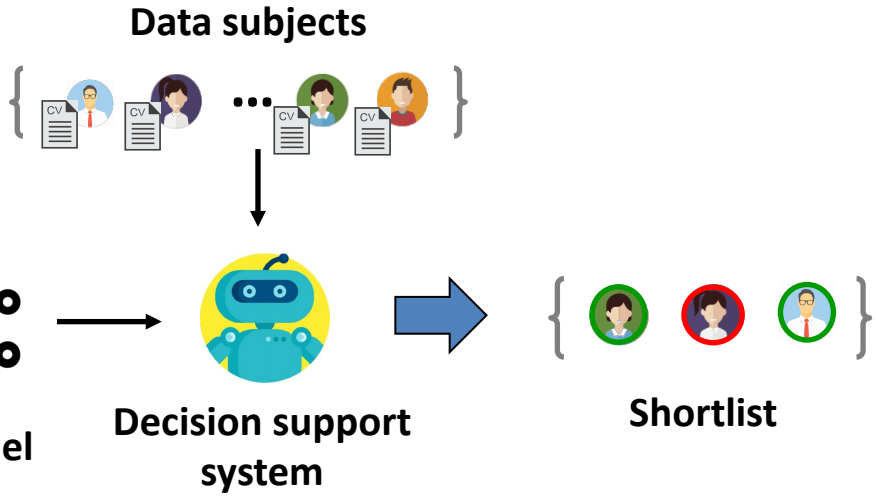
Artificial Intelligence to Improve Decision Making

AI promises a new **generation** of **decision support systems** in many **high-stakes domains**

Health

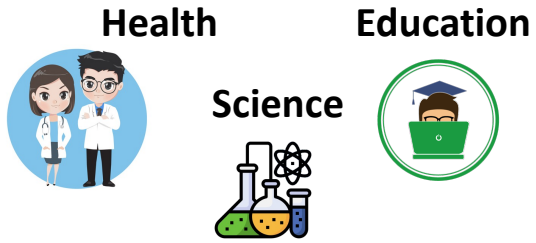
Science

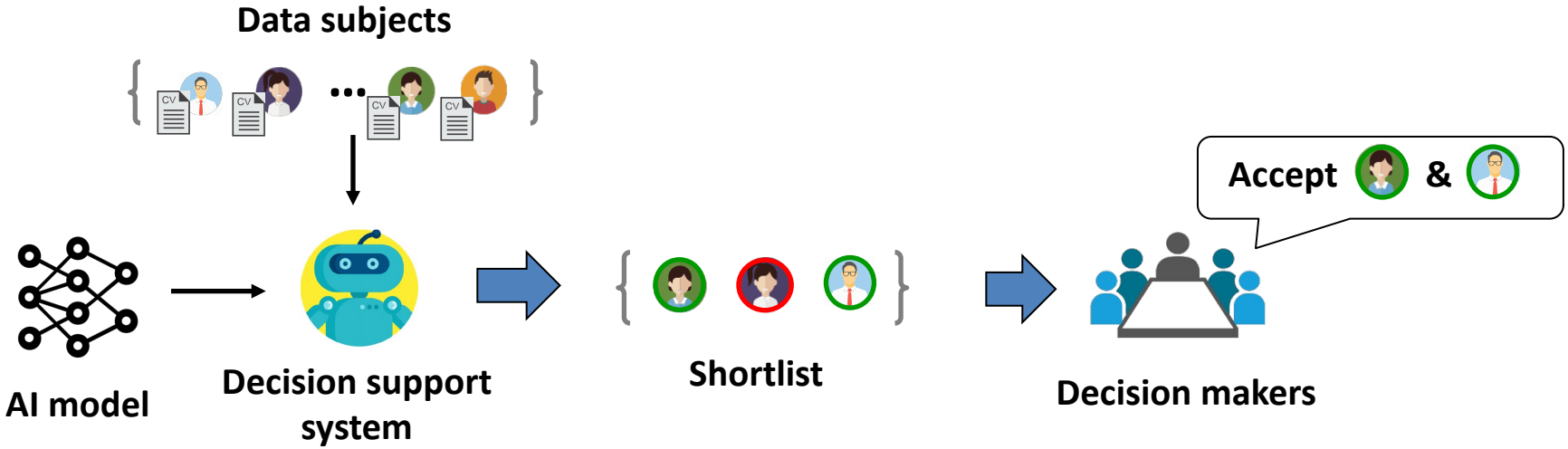
Education



Artificial Intelligence to Improve Decision Making

AI promises a new **generation** of **decision support systems** in many **high-stakes domains**



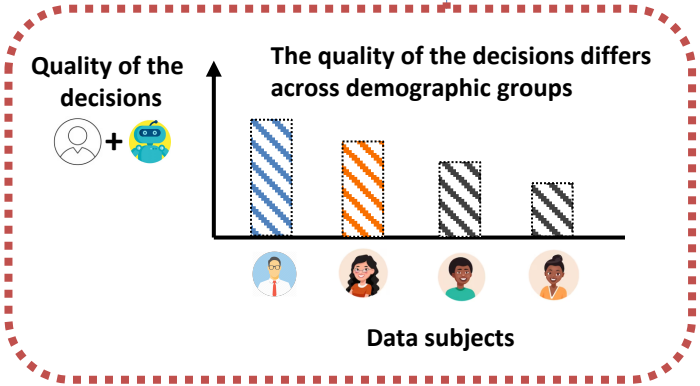


Diversity of Data Subjects, Decision Makers & AI Models

Decision support systems based on AI have not typically taken into account the diversity of data subjects, decision makers and AI models.

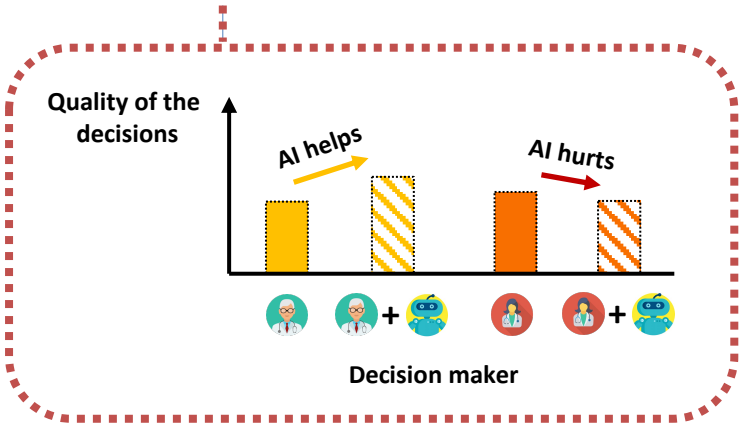
Diversity of Data Subjects, Decision Makers & AI Models

Decision support systems based on AI have not typically taken into account the diversity of data subjects, decision makers and AI models.



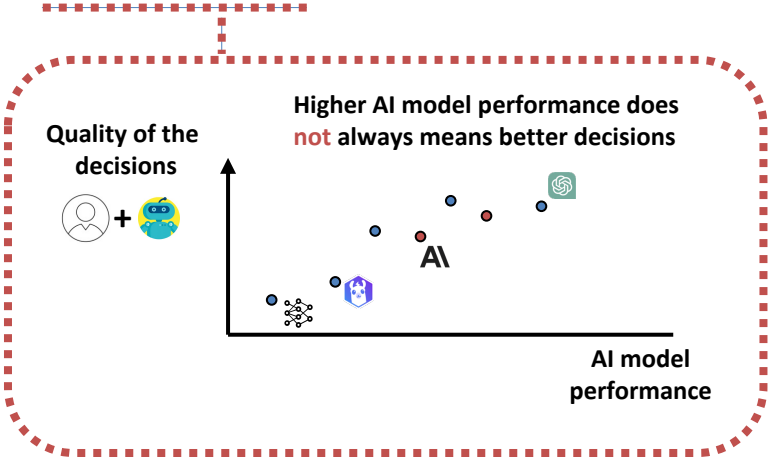
Diversity of Data Subjects, Decision Makers & AI Models

Decision support systems based on AI have not typically taken into account the diversity of data subjects, decision makers and AI models.



Diversity of Data Subjects, Decision Makers & AI Models

Decision support systems based on AI have not typically taken into account the diversity of data subjects, decision makers and AI models.



ERC StG project on Human-Centric AI

My ERC StG project HUMANML

is developing

AI-based decision support systems

for

all data subjects, decision makers & AI models

ERC StG project on Human-Centric AI

My ERC StG project HUMANML

is developing

AI-based decision support systems

for

all data subjects, decision makers & AI models

In the reminder of the talk, I will focus on one problem
regarding the diversity of data subjects

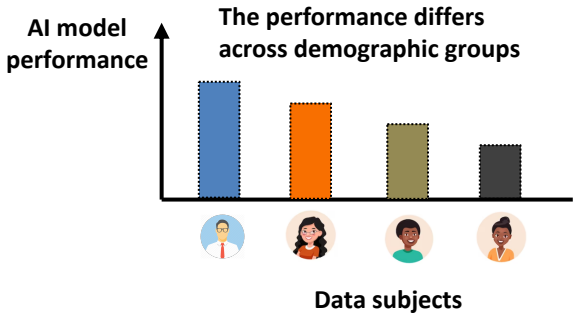
Fairness Across Groups

In recent years, the field of ethical AI has grown rapidly. There exist methods to ensure AI treats data subjects *across* different demographic groups *fairly*.

Fairness Across Groups

In recent years, the field of ethical AI has grown rapidly. There exist methods to ensure AI treats data subjects *across* different demographic groups *fairly*.

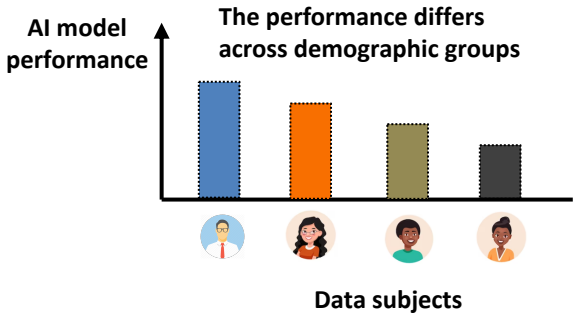
Before ethical AI



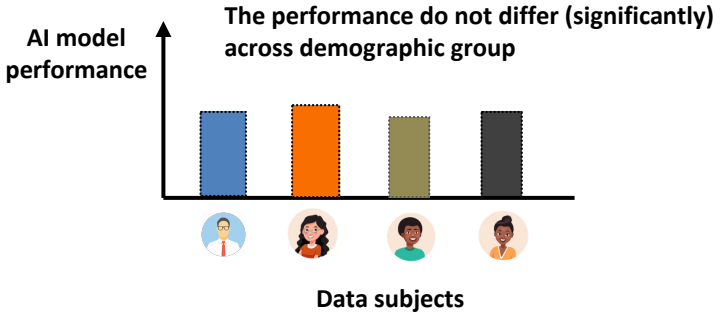
Fairness Across Groups

In recent years, the field of ethical AI has grown rapidly. There exist methods to ensure AI treats data subjects *across* different demographic groups *fairly*.

Before ethical AI



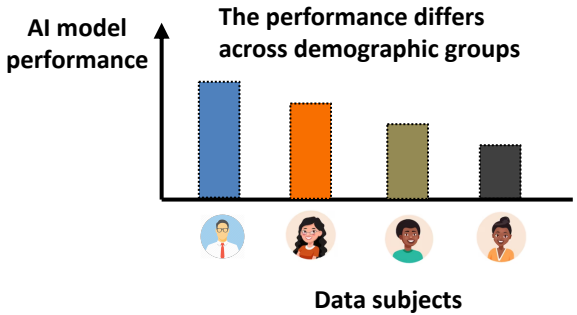
After ethical AI



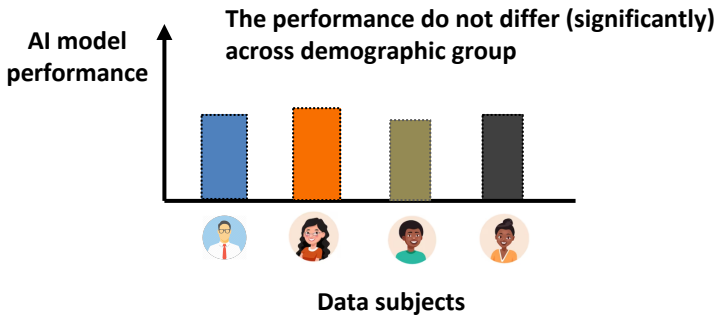
Fairness Across Groups

In recent years, the field of ethical AI has grown rapidly. There exist methods to ensure AI treats data subjects *across* different demographic groups *fairly*.

Before ethical AI



After ethical AI



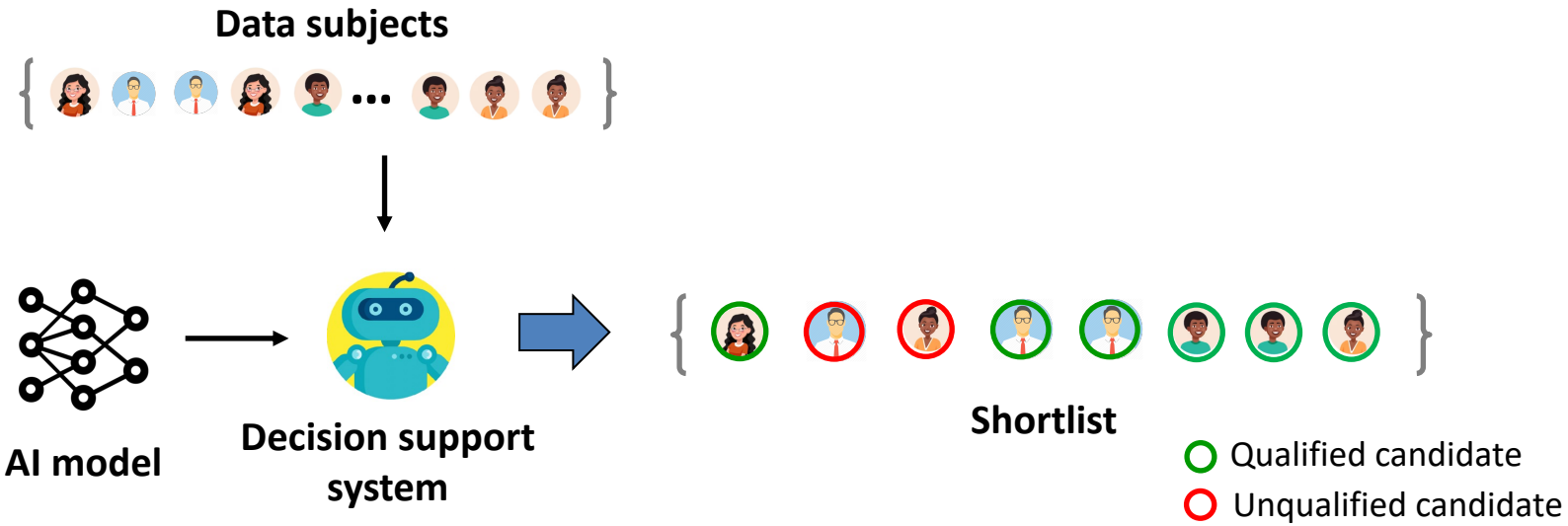
Are we done?

Unfairness Within Groups

However, existing methods do not ensure AI treats data subjects fairly *within* different demographic groups

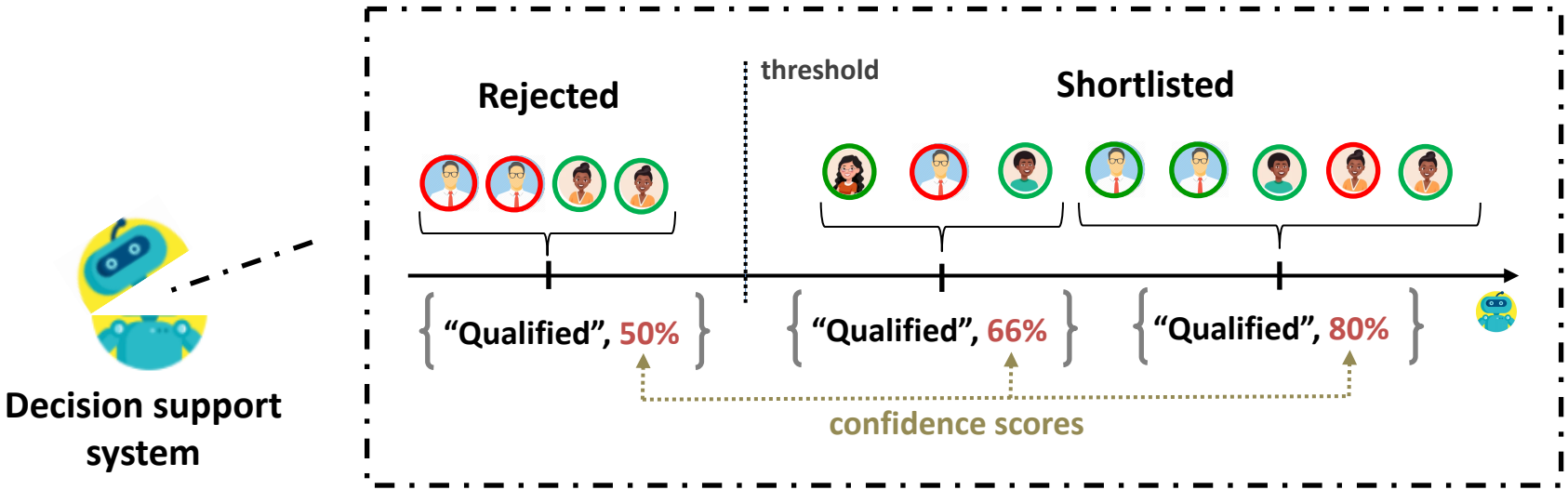
Unfairness Within Groups

However, existing methods do not ensure AI treats data subjects fairly *within* different demographic groups



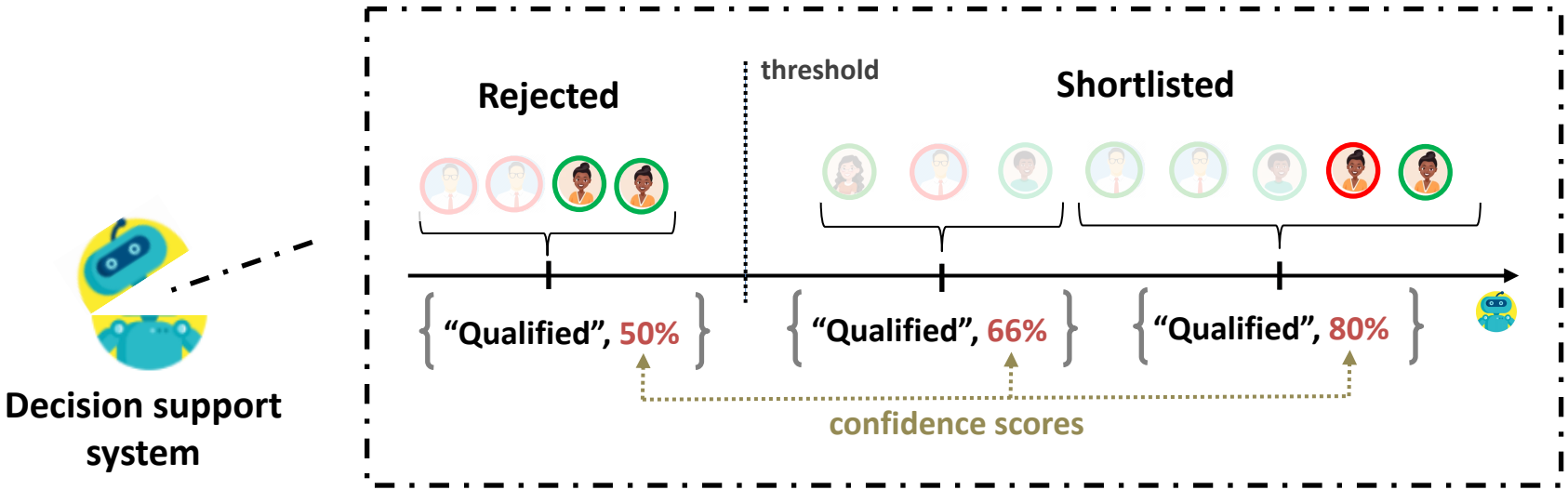
Unfairness Within Groups

However, existing methods do not ensure AI treats data subjects fairly *within* different demographic groups



Unfairness Within Groups

However, existing methods do not ensure AI treats data subjects fairly *within* different demographic groups



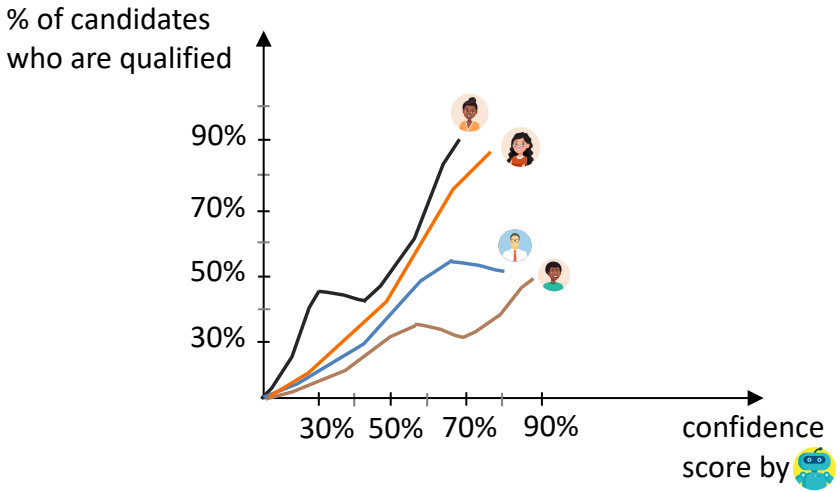
Within-group fairness

In my ERC project, we have developed a method to ensure AI satisfies *within-group* fairness, by construction

Within-group fairness

In my ERC project, we have developed a method to ensure AI satisfies *within-group* fairness, by construction

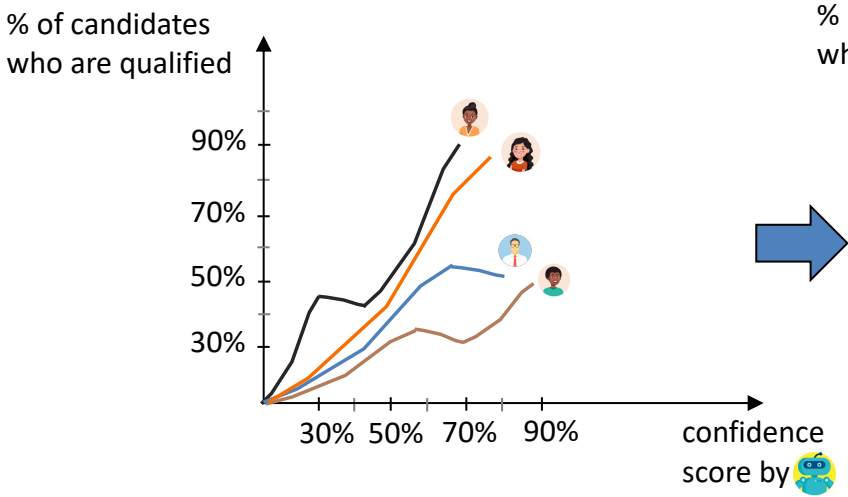
Before our project



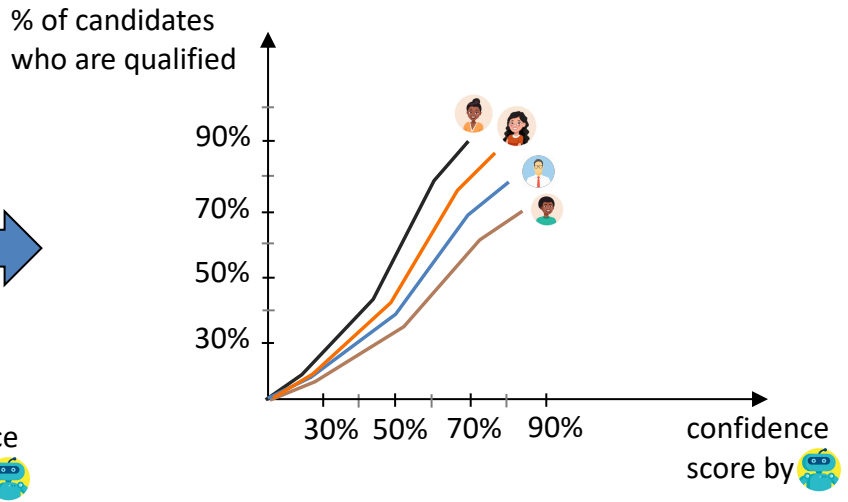
Within-group fairness

In my ERC project, we have developed a method to ensure AI satisfies *within-group* fairness, by construction

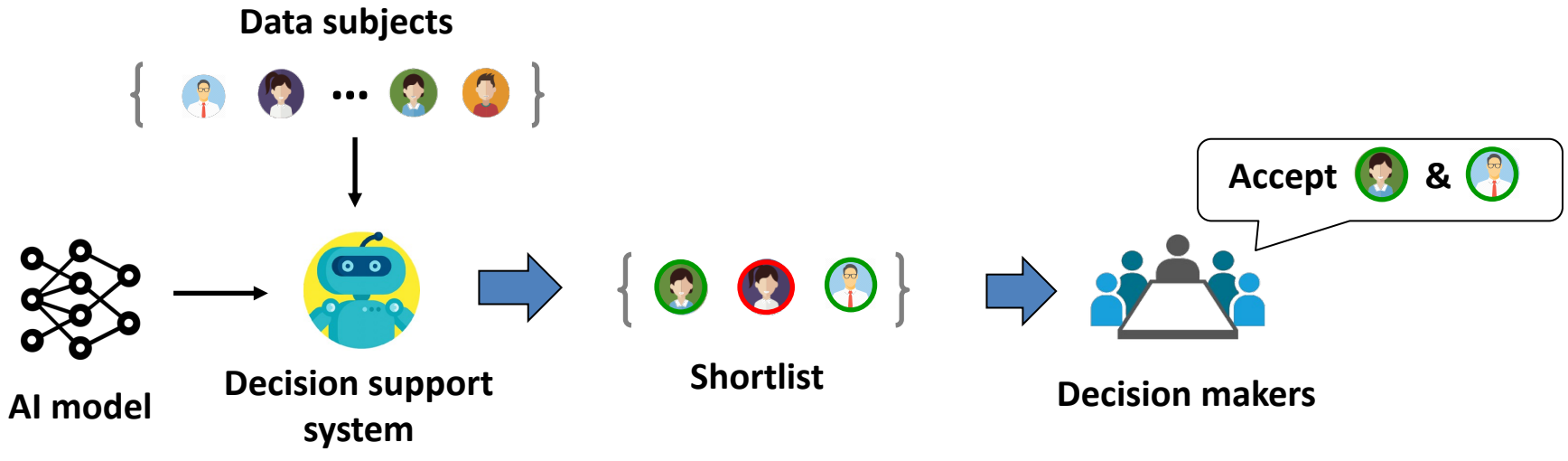
Before our project



After our project

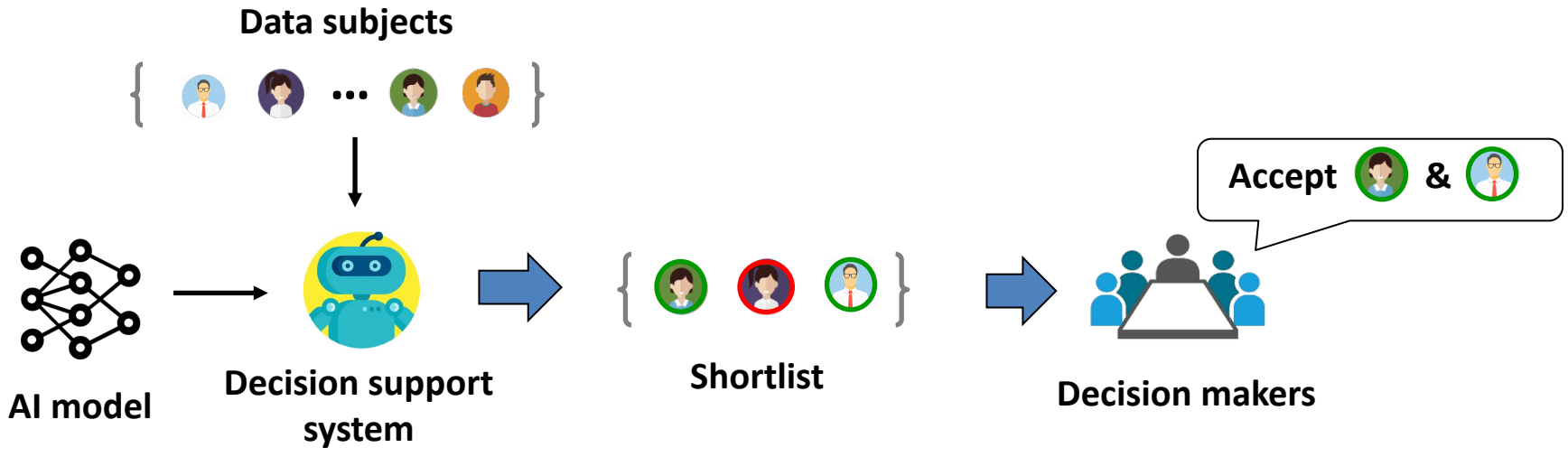


Validation and impact of within-group fairness



How did we validate it truly satisfies within-group fairness?

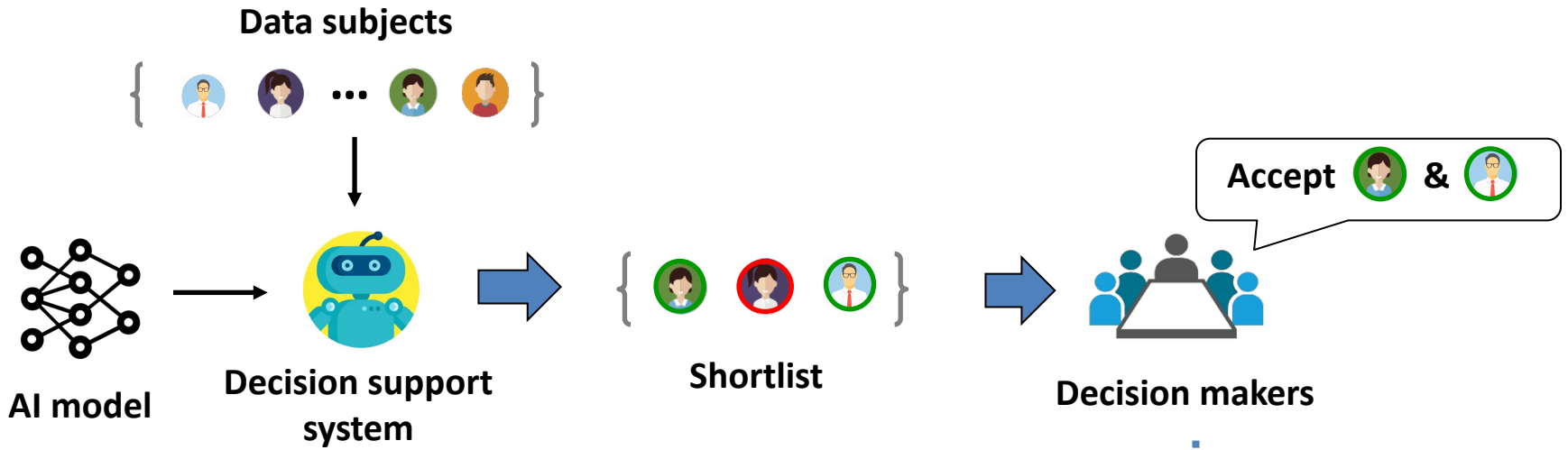
Validation and impact of within-group fairness



How did we validate it truly satisfies within-group fairness?

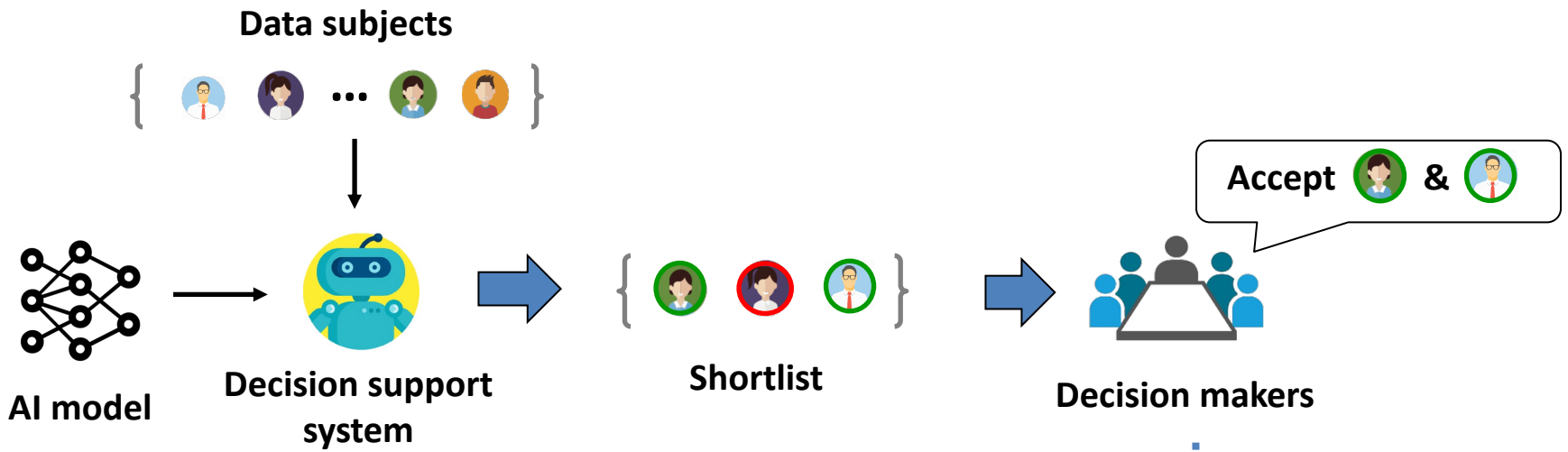
We simulated hundreds of screening processes using observational data about 3.2 million people from the US Census

Validation and impact of within-group fairness



What about the impact of within-group fairness on the quality of the decisions?

Validation and impact of within-group fairness



What about the impact of within-group fairness on the quality of the decisions?

Very challenging, if not impossible, to assess impact with observational data. Need to run human subject studies!

Thanks!

On the Within-Group Discrimination of Screening Classifiers

ICML 2023

Nastaran

Stratis

Learn more about our research at
learning.mpi-sws.org