Fairwashing: the risk of rationalization

Ulrich Aïvodji

UQÀM

Sébastien Gambs

UQÀM

Hiromi Arai

AIP

Olivier Fortineau

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIE
PARIS INSTITUTE OF TECHNOLOGY

Satoshi Hara

大阪大学
OSAKA UNIVERSITY

Alain Tapp

Université de Montréal

High social demands on ethically aligned AI

IEEE
Advancing Technology for Humanity

ETHICALLY ALIGNED DESIGN
First Edition Overview
A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

Ethically Aligned Design, First Edition Overview

Praise for *Ethically Aligned Design*

"To create and foster trust between humans and machines, you must understand the ethical resources and standards available for reference during the designing, building, and maintenance of AI. The large-scale focus on AI ethics by groups like The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems... should be mirrored in businesses and working groups of all sizes."

- IBM: Everyday Ethics for Artificial Intelligence; A practical guide for designers & developers

"With...Ethically Aligned Design...there could be every chance companies will move beyond good intentions and into standardized practices that factor human well-being into design."

- Mashable

"This document, which will continue to change and grow with the times, encourages those pushing technology forward to consider how such progress might interfere with ethical concerns. The nature of the publication, that it is 'living', shows how well its creators at IEEE understand the ever-changing nature of the AI field."

- Futurism

"How often do you come across a paper from the tech community that says, 'Human well-being is the highest virtue for a society, and human flourishing begins with conscious contemplation'? That is the tech community that I recognize, a community of intelligent and articulate people with a genuine desire to make the world a better place. A community for which I am proud to belong."

- Australia's Chief Scientist, Dr. Alan Finkel, IEEE Sections Congress, August 2017

"As the tech world pushes forward with the development of artificial intelligence, The Institute of Electrical and Electronics Engineers (IEEE) is asking everyone to pause and consider the ethical ramifications."

- ZDNet

"The organization has published a framework document it's hoping will guide the industry toward the light—and help technologists build benevolent and beneficial autonomous systems, rather than thinking that ethics is not something they need to be worrying about."

- TechCrunch

High social demands on ethically aligned AI

II. General Principles

The ethical and values-based design, development, and implementation of autonomous and intelligent systems should be guided by the following General Principles:

1. **Human Rights**
   A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.

2. **Well-being**
   A/IS creators shall adopt increased human well-being as a primary success criterion for development.

3. **Data Agency**
   A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to have control over their identity.

4. **Effectiveness**
   A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS.

5. **Transparency**
   The basis of a particular A/IS decision should always be discoverable.

6. **Accountability**
   A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.

7. **Awareness of Misuse**
   A/IS creators shall guard against all potential misuses and risks of A/IS in operation.

8. **Competence**
   A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.

High social demands on ethically aligned AI

1. Human Rights
A/IS shall be created and operated to respect, promote, and protect internationally recognized human rights.

We expect the decisions made by AI to be fair.
No discrimination over gender, race, …

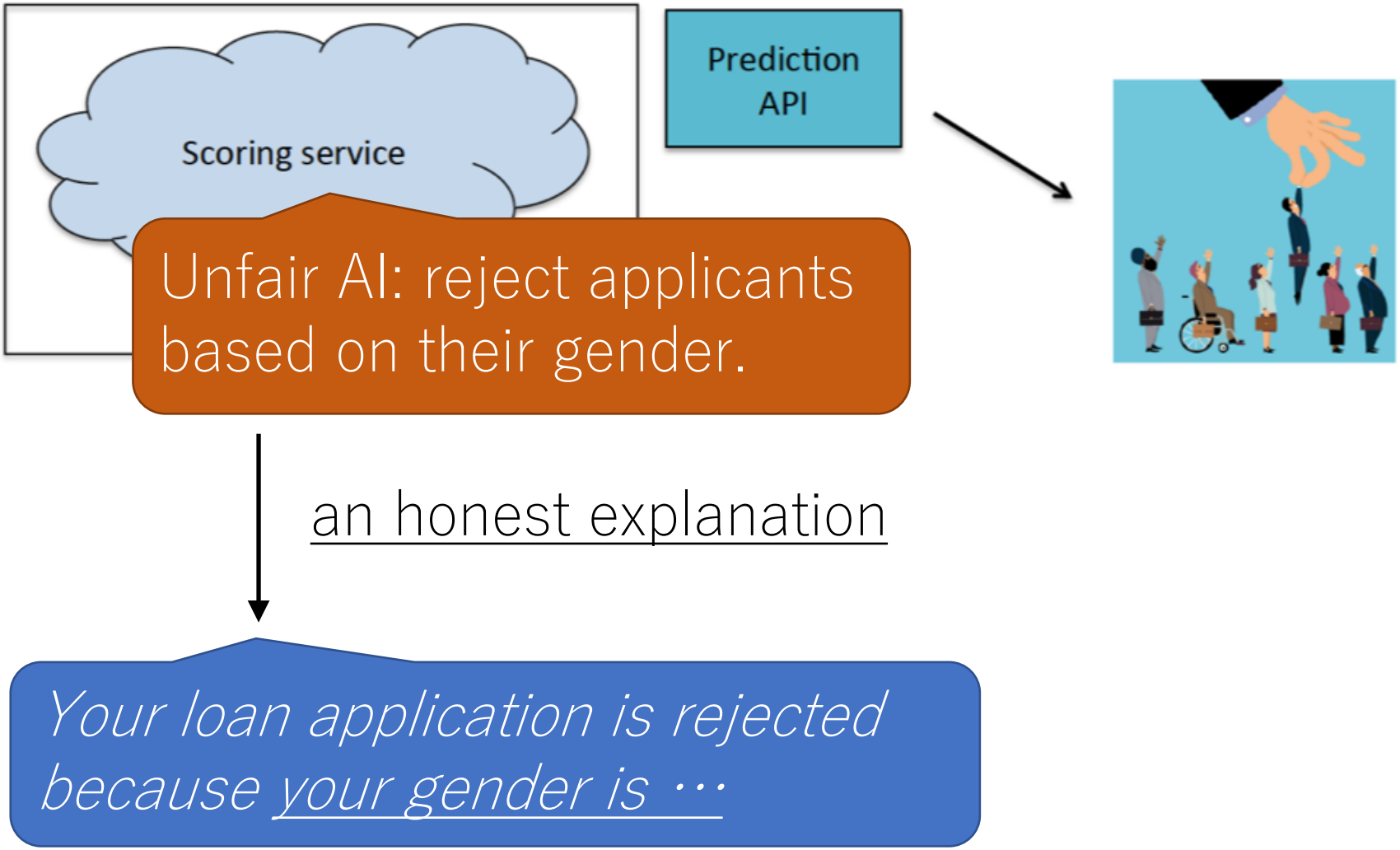
5. Transparency
The basis of a particular A/IS decision should always be discoverable.

6. Accountability
A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.

We expect AI to explain the reasons of the decisions.
Your loan application is rejected because …

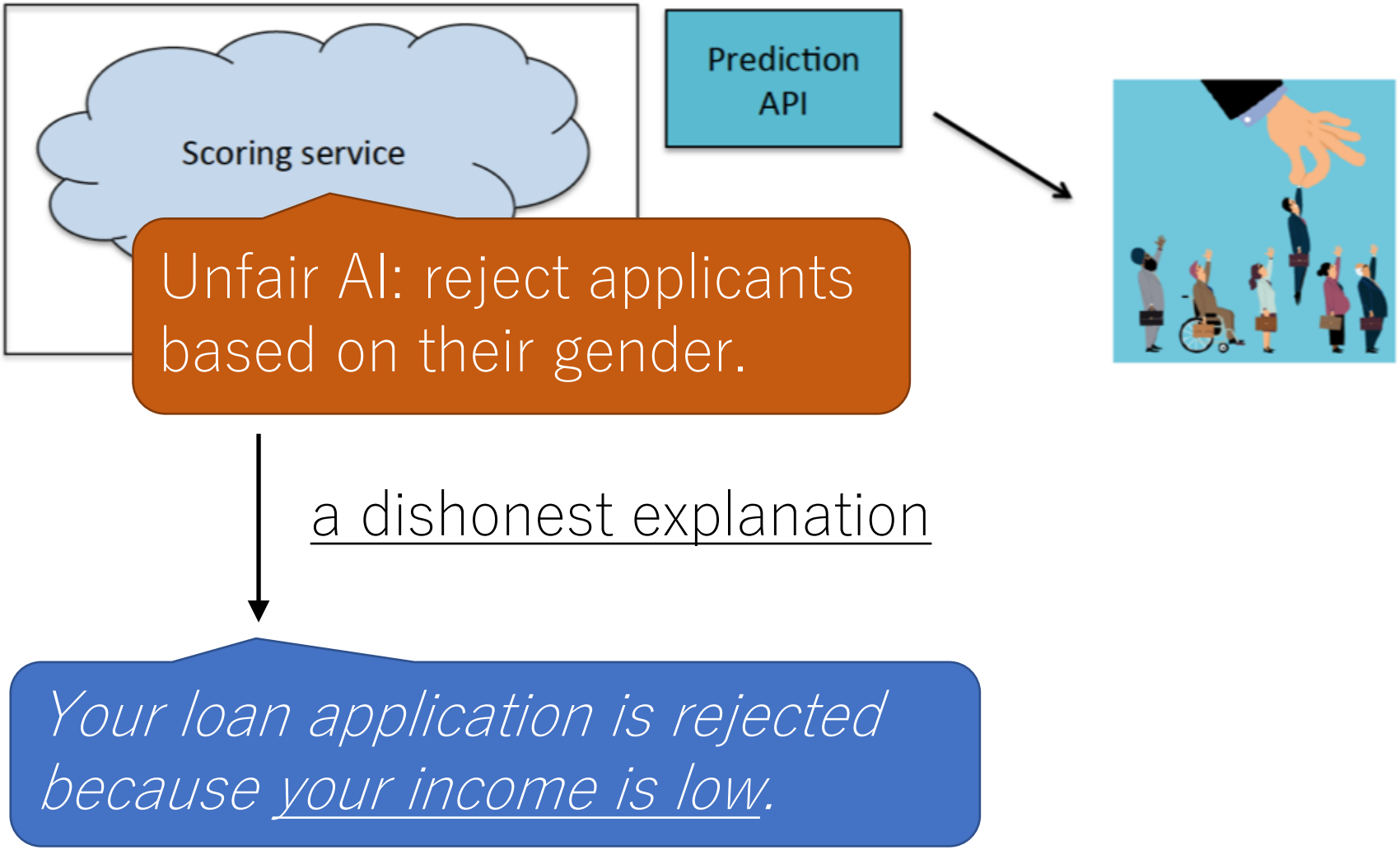
A pitfall: The risk of “Fairwashing”

- Explaining fairness



A pitfall: The risk of “Fairwashing”

- Explaining fairness



A pitfall: The risk of “Fairwashing”

- Explaining fairness

“Fairwashing”

Malicious decision-makers can disclose a *fake explanation* to rationalize their unfair decisions.

A pitfall: The risk of “Fairwashing”

- Explaining fairness

“Fairwashing”

Malicious decision-makers can disclose a *fake explanation* to rationalize their unfair decisions.

This Study: LaundryML
Possible to systematically generate fake explanations.

Raise the awareness of the risk of “Fairwashing”.

LaundryML
Systematically generating fake explanations

The idea

Generate many explanations,
and pick one that is useful for “Fairwashing”.

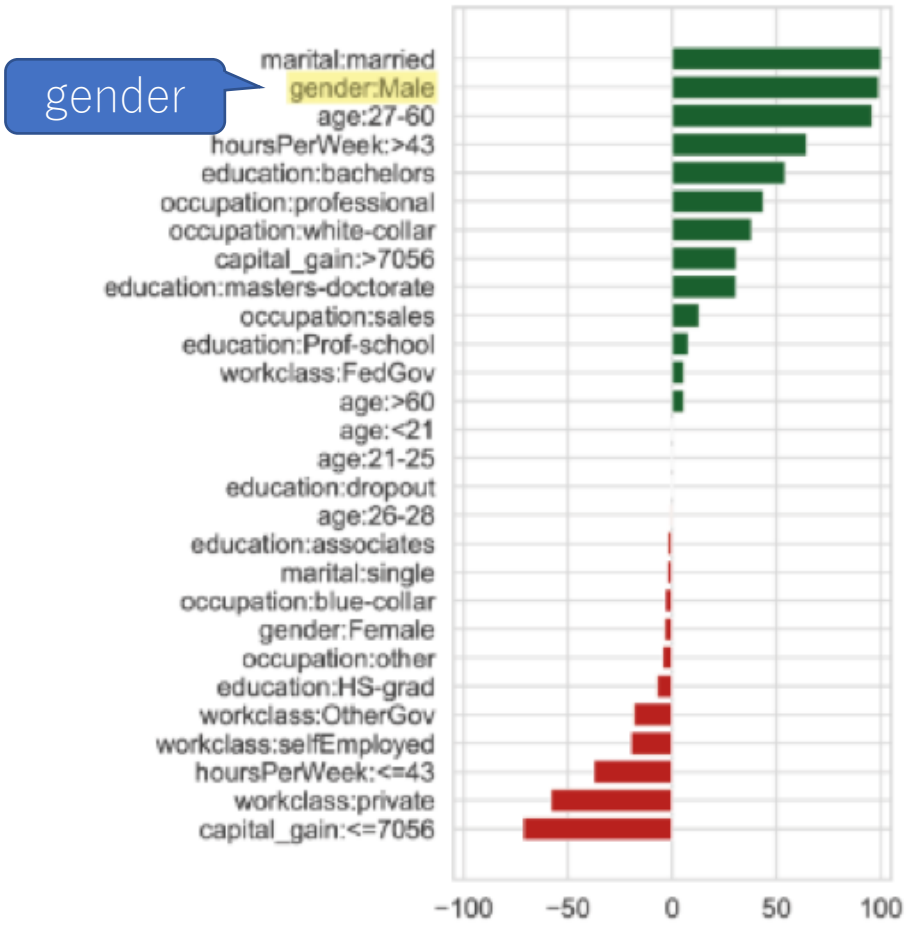
- many explanations
  - Use “Model Enumeration” [Hara & Maehara’17; Hara & Ishihata’18]
  - Enumerate explanation models.

- pick one
  - Use fairness metrices such as demographic parity (DP).
  - Pick an explanation most faithful to the model, with DP less than a threshold.

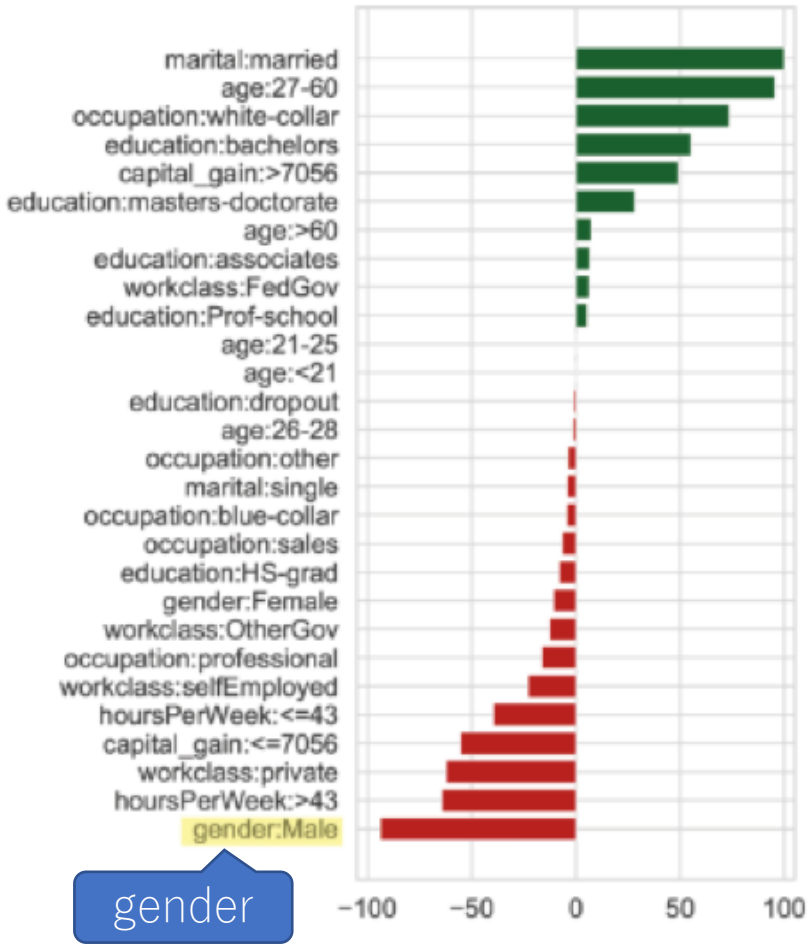
Result

- "Fairwashing" for decisions on Adult dataset
  - Feature importance by FairML on "gender" has dropped.

A naïve explanation



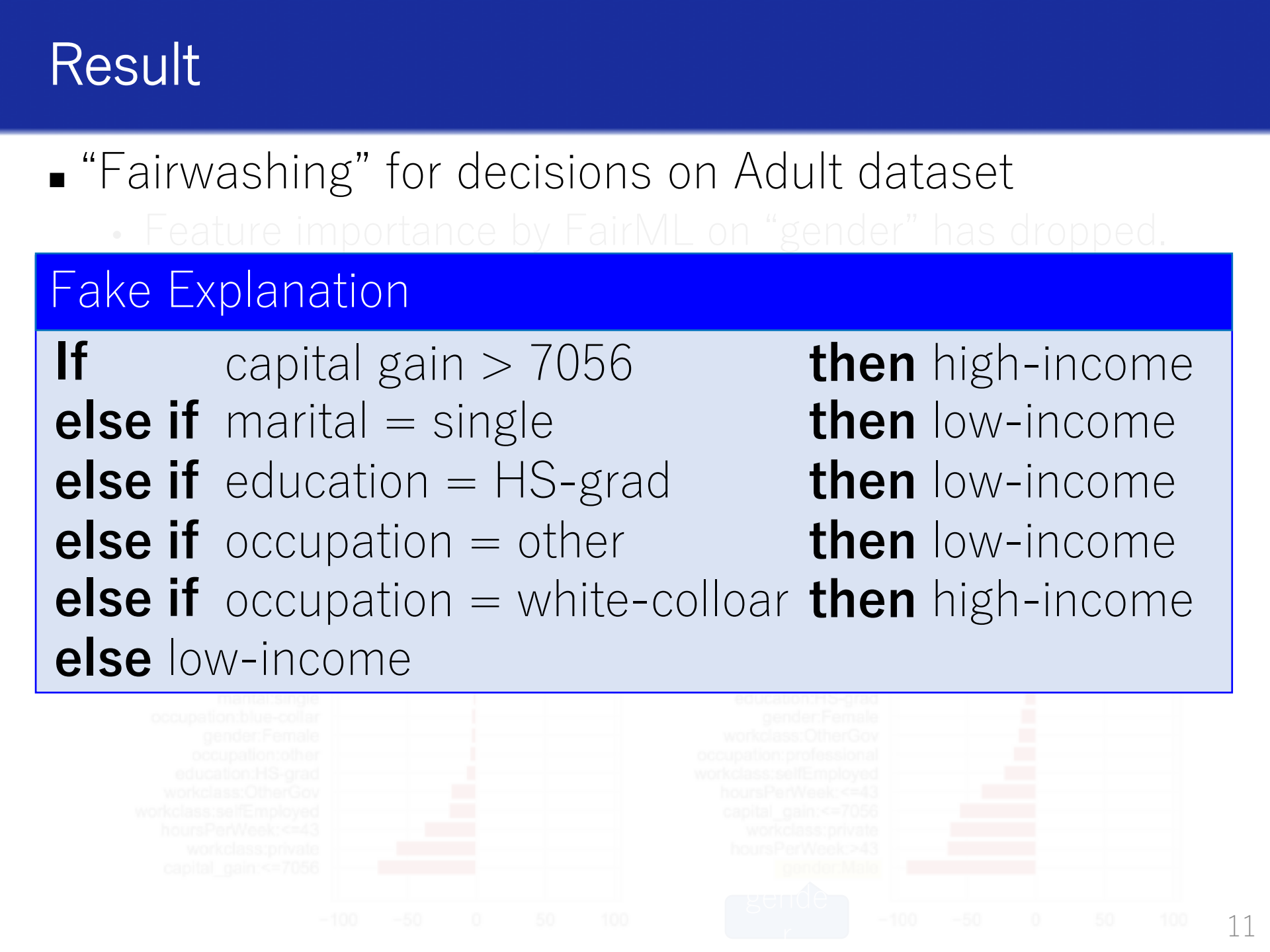
A fake explanation



Result

- "Fairwashing" for decisions on Adult dataset
  - Feature importance by FairML on "gender" has dropped.

Fake Explanation		
If	capital gain > 7056	**then** high-income
else if	marital = single	**then** low-income
else if	education = HS-grad	**then** low-income
else if	occupation = other	**then** low-income
else if	occupation = white-colloar	**then** high-income
else low-income		



Summary

"Fairwashing"

Malicious decision-makers can disclose a *fake explanation* to rationalize their unfair decisions.

- We raise the awareness of the risk of "Fairwashing".
- With LaundryML, we can systematically generate fake explanations.
- Question: How can we avoid "Fairwashing"?