

# Start Using Justifications When Explaining AI Systems to Decision Subjects

---

**Klára Kolářová, ETH Zürich**

[kkolarova@student.ethz.ch](mailto:kkolarova@student.ethz.ch)

**Timothée Schmude, University of Vienna**

[timothee.schmude@univie.ac.at](mailto:timothee.schmude@univie.ac.at)



[To the paper](#)

# Position

***People who are affected by automated decisions  
not only need explanations of how decisions are made,  
but also justifications of why they are legitimate.***

# What are automated decisions?

- Made by systems that assist or replace human decision-making
- Increasingly embedded in institutional procedures
- Often operate as “black boxes”

**EXAM SHAMBLES** Fears of GCSE results meltdown as grades will be awarded using algorithm behind A Level fiasco

Ben Hill

Published: 00:58, 16 Aug 2020

The Sun 08/20

UK creating ‘murder prediction’ tool to identify people most likely to kill

Exclusive: Algorithms allegedly being used to study data of thousands of people, in project critics say is ‘chilling and dystopian’

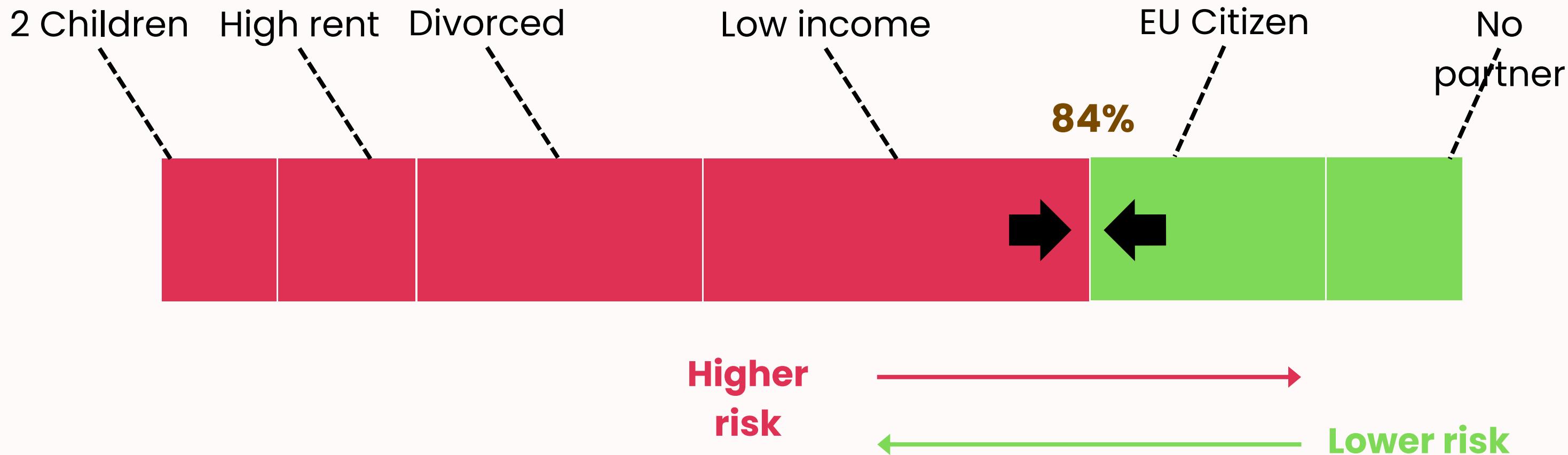
The Guardian 04/25

An algorithm that screens for child neglect raises concerns

AP 04/22

# What explanations usually do

- Help humans understand how a model reaches its outputs → **descriptive**
- **Best approximations**, not ground truth



# What explanations should also do

- *Justify the goals and values that guide design and deployment* → **normative**
- **Every automated system embeds human intentions:**
  - automates **human-defined rules**, or
  - learns from **past human decisions**, or
  - optimises a **human-chosen target** using **human-chosen data**

**These are explainable!**

# Why do we need justifications?

- 1. To address decision subjects' epistemic needs**
- 2. To assign accountability throughout system lifecycle**
- 3. To support decision subjects in accepting or contesting decisions**



- *What will happen to me now?*
- *Is this fair?*
- *How can I change this?*
- *Who can I contact?*

# Designing justifications

***Justifications describe the goals and values that guide design and deployment of an automated system***

Good justifications are

- 1. Normative** – focus on values and intentions rather than mechanisms
- 2. Argumentative** – provide multiple viewpoints
- 3. Challengeable** – invite opposition
- 4. Relational** – adjust to the knowledge of the recipient

# Key takeaways

- **Decision subjects need justifications, not only descriptive explanations**
- **Justifications:**
  - **address decision subjects' epistemic needs**
  - **support accountability**
  - **support acceptance or contestation**
- **Good justifications are normative, argumentative, challengeable, relational**

# Thank you!

---

**Klára Kolářová, ETH Zürich**

[kkolarova@student.ethz.ch](mailto:kkolarova@student.ethz.ch)

**Timothée Schmude, University of Vienna**

[timothee.schmude@univie.ac.at](mailto:timothee.schmude@univie.ac.at)



To the paper