Bridging the Transparency Gap: What Can Explainable AI Learn From the AI Act?

Balint Gyevnar, Nick Ferguson, Burkhard Schafer

THE UNIVERSITY *of* EDINBURGH
informatics

Transparency Gap

What is transparency?

Artificial Intelligence Act: it is a <u>means</u> to achieve wider values

Protect human rights   *Sustainable innovation*   *Accountability*

Explainable AI: it is limited to be an <u>end</u> in itself

Algorithmic understanding   *Debugging*   *Build trust*

The Issue

No mutual conceptual understanding for:

- XAI to consider its social impact;
- Legislators to assess what is technically feasible.

Scope of Transparency

Wider view considers social motivations and consequences

Transparency required to an *appropriate level*

AIA ➞ **XAI**

"Mark your own homework" approach

(Over-)reliance on standard setting bodies

Extend motivation beyond direct effects of methods

Involve humans in the evaluation of explanations

Outline clear guidelines for external bodies

Define what counts as *substantial modification*

Conformity Assessment

Legal Status

Unclear whether XAI tools are considered AI under the AIA

What *outputs* of an AI system are covered under AIA?

AIA ➞ **XAI**

Detailed and strict data governance requirements:
- Data source, composition;
- Bias, purpose;
- Relevancy, etc…

XAI tools should be considered the same with as the AI system

Define clearly what outputs the XAI tools produces

Explainable dataset documentation (e.g., datasheets)

XAI for exploring and explaining properties of data

Explainable Data