Bridging the Transparency Gap

What can Explainable AI learn from the AI Act?

Bálint Gyevnár · Nick Ferguson · Burkhard Schafer

THE UNIVERSITY of EDINBURGH
UKRI Centre for Doctoral Training
in Natural Language Processing

Read the paper here.

Overview and contribution.

The EU's proposed **Artificial Intelligence Act** sets out detailed **transparency requirements** for AI systems.

The Act views transparency as a **means to support wider values**, while explainable AI research views transparency as **an end in itself**.

Mutual conceptual understanding in necessary for XAI researchers to consider the social impact of their work, and for legislators to assess what is feasible.

We address this **transparency gap** on four fronts:

1. The **scope** of Transparency.
2. The **legal status** of XAI.
3. Requirements for **conformity assessment.**
4. Building the role of **explainable datasets.**

What does the Act say about transparency?

User-empowering transparency

- User-empowering transparency is concerned with **enabling users** to **understand systems' output**.
- It also governs **instructions for use**, requiring them to be *concise, clear, correct and complete*; and contain information about *characteristics, capabilities, and limitations of performance*.
- This is to enable effective **human oversight**.

Compliance-oriented transparency

- Compliance-oriented transparency are focused on transparency through extensive **technical documentation**.
- Documentation must include information on *design* and *classification choices*
- Other compliance-oriented requirements include **risk** and **quality management** systems, and a **record keeping** system.

Four recommendations.

Key observations → Our recommendations

The Scope of Transparency

Different interpretations of *transparency* affect the form of transparent AI.

The Act says transparency is an **overarching property** of the AI system. XAI views transparency in an **algorithmic** sense.

Requirements for an *appropriate level* of transparency is ambiguous, and *full understanding* of capacities and limitations is infeasible.

Alternatively, XAI views on transparency lack awareness of the wider **societal context**.

Requirements on appropriate levels of transparency should be relative to **limitations** and **intended purpose**.

XAI should remain in pursuit of social and human-centred XAI, looking towards the Act's requirements for guidance.

Legal Definition of XAI

The nature of regulation requires some flexibility, yet unfit definitions of key concepts creates confusion.

For one, it is unclear whether XAI systems should be **treated as AI systems themselves**.

Additionally, what is considered the **output** of an AI system requires deeper thought. The Act refers to *content* and *predictions*, yet XAI systems often rely on 'internal outputs', such as model weights.

We recommend that XAI systems are **not** considered AI systems to avoid regulatory issues.

What constitutes an *output* of a system should not inhibit the ways in which XAI systems can be implemented.

Conformity Assessment

The relationship between XAI and AI systems adds complexity to conformity assessment.

The *mark your own homework* approach lacks involvement of external organisations, such as notified bodies or standards-setting organisations.

Dominant involvement of standards-setting bodies may **reduce protection of basic rights** by over-emphasising **compliance-oriented transparency**.

The involvement of external bodies should be clarified to make routes to proper conformity assessment clear.

Defining what constitutes a *substantial modification* of a system will affect when recertification is required.

Explainable Datasets

The potential for the use of data to promote explainability is underutilised in the Act.

The Act places requirements on the quality of *training*, *validation*, and *testing* data. Yet, this paradigm of data use is only appropriate for some AI systems.

The extent to which XAI is leveraged to explain datasets is not fully realised in the Act.

The Act should better acknowledge **other data paradigms**, e.g., for reinforcement learning and planning.

The Act could exploit the potential of XAI for data to give subjects more information about it's inherent biases.

UK Research and Innovation
UKRI CENTRE FOR DOCTORAL TRAINING
THE UNIVERSITY of EDINBURGH informatics