Text, logo

Description automatically generatedLogo

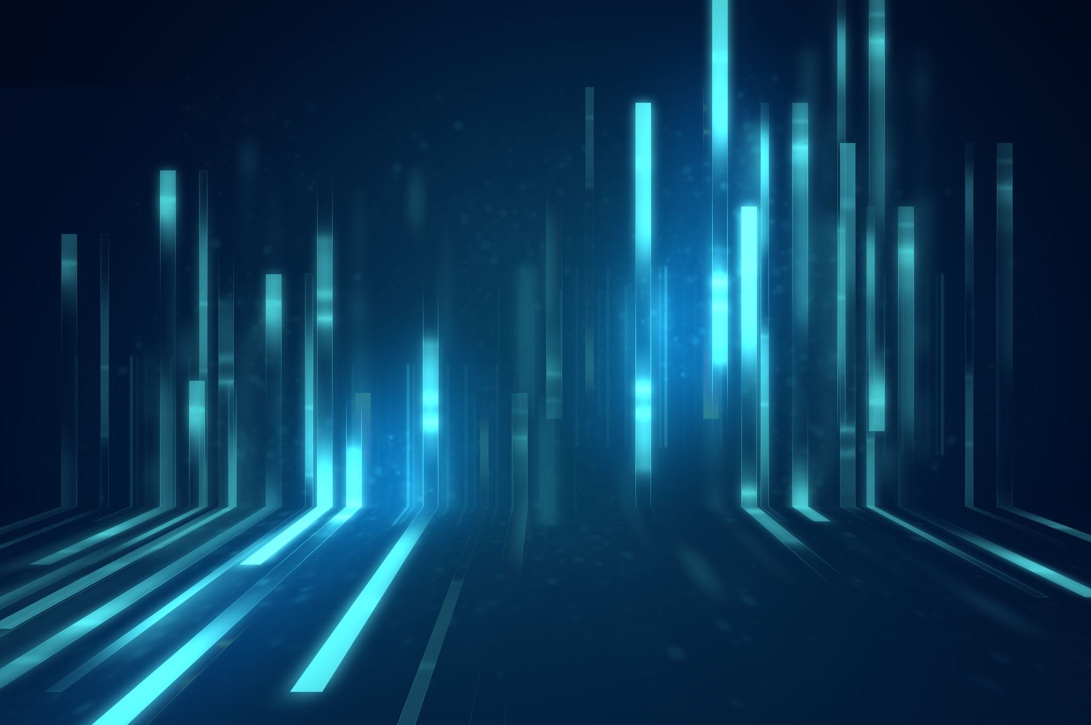
Description automatically generated with medium confidence

­­­

Transparency Intermediate

Module 2

**Explainability**



# Introduction

A closely related value to transparency is explainability. Explainable AI (XAI), for example, is artificial intelligence which outputs can be understood by humans. In this module we will

* Examine the relation between transparency and explainability
* Discuss the difference between backend explainability and frontend explainability
* Explain different interpretations of explainability (the process of understanding vs the quality of understanding)
* Identify cases of insufficient explainability and transparency so that they can be addressed
* Discuss approaches for evaluation of explanation

­­­

# Reading

When — and Why — You Should Explain How Your AI Works

<https://hbr.org/2022/08/when-and-why-you-should-explain-how-your-ai-works>

Reijers, W., Wright, D., Brey, P., Weber, K., Rodrigues, R., O’Sullivan, D., & Gordijn, B. (2018). Methods for Practising Ethics in Research and Innovation: A Literature Review, Critical Analysis and Recommendations. Science and Engineering Ethics, 24(5), 1437-1481. <https://doi.org/10.1007/s11948-017-9961-8>

Turilli, M., & Floridi, L. (2009). The ethics of information transparency. Ethics and Information Technology, 11(2), 105-112. <https://doi.org/10.1007/s10676-009-9187-9>

Worthington, R. (1982). The Social Control of Technology. By David Collingridge. (New York: St. Martin's Press, 1980. Pp. i + 200. $22.50.). American Political Science Review, 76(1), 134-135. <https://doi.org/10.2307/1960465>

OECD AI Principles overview

<https://oecd.ai/en/ai-principles>

It’s better to understand something than to know it

<https://qz.com/1123896/its-better-to-understand-something-than-to-know-it>

Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue

<https://arxiv.org/pdf/2108.03012.pdf>

­­­



**Pre-reading**

Explaining AI – How to explain the unexplainable?

<https://www.hiig.de/en/explaining-ai-explain-the-unexplainable>

Explanation in Artificial Intelligence: Insights from the Social Sciences

<https://arxiv.org/abs/1706.07269>

**Reflection sheet**

1) Why is XAI considered to be valuable? Are there any disadvantages to doing XAI?

2) List and describe 3 of the technical methods described in the module for doing XAI.

3) Is explainability a subcategory of transparency or its own category? Explain your answer.

4) Provide 2 reasons from the module why XAI may not be necessary at all.

**Tasks**

**Task 1**

* Make notes to give a five-minute talk explaining to doctors the advantages and disadvantages of using XAI for their diagnostic tools, ending with practical steps they can take to ready themselves for using AI tools.

**Task 2**

* Read the following article and think about the example of the parolee at the top of the page

<https://itrexgroup.com/blog/explainable-ai-principles-classification-examples>

* Questions
  + How could this have been avoided?
  + Using the principles learned in this module, what could you have done from a technical perspective to help achieve a different outcome?

**Task 3**

Group Exercise

* From the previous article about the parolee in NY, imagine the company that built the AI used to adjudicate the man’s parole is being taken to court. Appoint a third of the group to argue for the company, a third of the class to argue against, and the final third to act as the court, marshalling those arguments into a final judgment.
* At the end, ask
  + Is the class satisfied with the judgment?
  + Do they think that the judgment is realistic and would happen today?
  + Is a company responsible for poor explainability even for examples it does not foresee?

**Self-assessment pass/fail questions**

1) What is one of the differences between knowing and understanding?

a. Interrelatedness of the subject matter

b. How convinced you are that you are correct

c. If you are able to tell other people about it

2) Are interpretability and explainability the same thing?

a. Yes

b. No

c. Sometimes

3) How can we best explain algorithms?

a. By isolated bits of information

b. Using generalisable models

c. Aligning explanations with the decision processes of the end users

4) Are there metrics available for evaluating the quality of an explanation?

a. Yes, and they are simple to apply

b. No, we cannot know whether an explanation is good or bad.

c. Yes, but the field is still new meaning we should be cautious regarding our conclusions.

**Answers**

Qs 1) D, 2) B, 3) A, 4) B, 5) B, 6) D, 7) A Q Qs