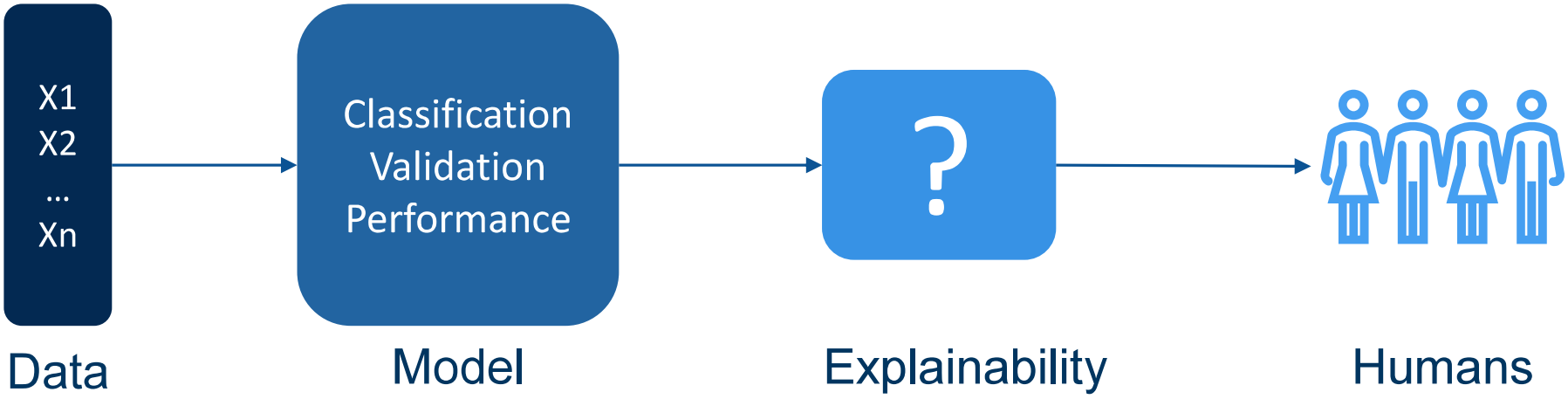
‘Explainability’ in Healthcare Applications

Dr. Fani Deligianni,
fani.deligianni@glasgow.ac.uk
Lecturer (Assistant Professor)
Lead of the Computing Technologies for Healthcare Theme
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

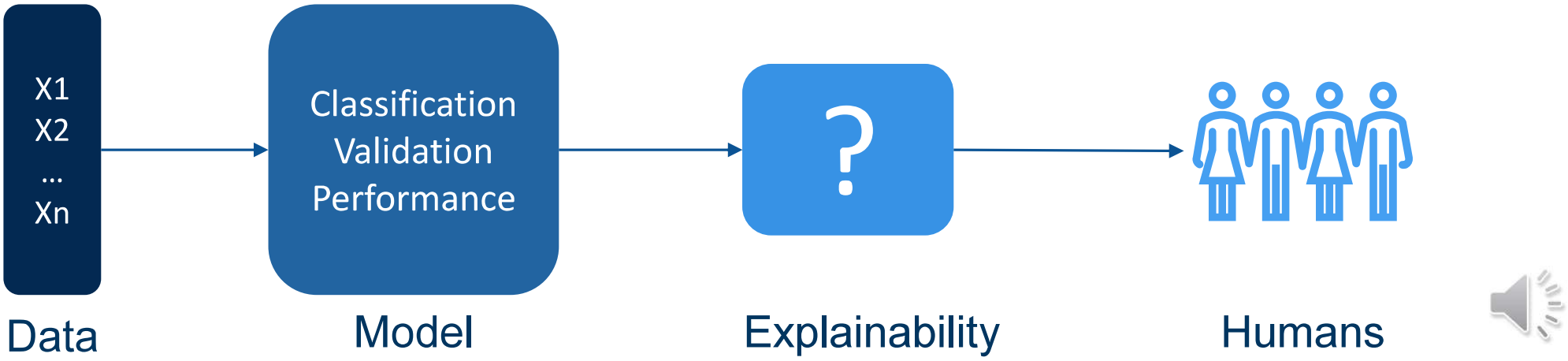
Explainability in Healthcare

- Stiglic et al. (2020) describes the need for explainability of ML models in healthcare.
- The need for explainable 'black-box' models is important as it will allow the clinicians to make **decisions driven by data.**
- This would in turn help patients receive **personalized high-quality care.**
- It will improve efficiency in **recognizing specific abnormalities** help accelerate and optimize treatment.



Lack of Explanation is Harmful

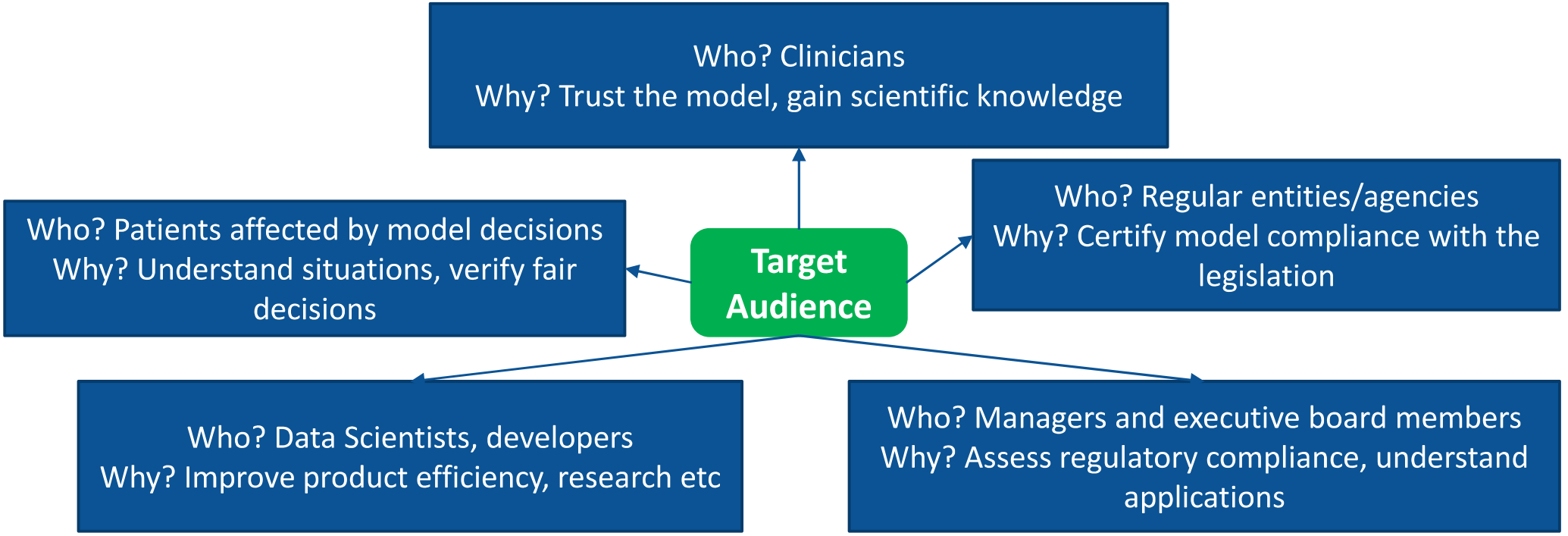
- There are many ways that AI systems can behave unfairly.
- For example, AI can impact the quality of service which we use, again whether system works as well for one person as it does for another.
- AI can also impact allocation; AI systems can extend or withhold opportunities, resources or information to specific groups of people.



Importance of ‘Explainability’

- Explainability is required to ensure impartial decision-making process
  - Detect biases, ensure fairness
- Explainability ensures that only meaningful variables infer the output
  - Explain the decision-making process
  - Fundamental human right to know why
- Explainability ensures robustness of the results

Target Audience of Explainability



‘Explainability’ - Goals

- Trustworthiness
- Causality
- Transferability/Generalisation
- Informativeness
- Confidence
- Fairness
- Accessibility
- Interactivity
- Privacy awareness

Summary

- Machine learning models cannot be adapted in healthcare unless, they are explainable
- ‘Explainability’ is a key requirement for a clinical decision making system because it relates with trustworthiness
- The relationship between ‘Explainability’ and ‘Trustworthiness’ is not reciprocal

References

- Arrieta et al. ‘Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI’, Information Fusion, 2020.
- Molnar ‘Interpretable Machine Learning - A Guide for Making Black Box Models Explainable’, 2021. https://christophm.github.io/interpretable-ml-book/
- Stiglic et al. ‘Interpretability of machine learning-based prediction models in healthcare’, 2020.