Machine Learning and AI are considered by many as techniques free of personal judgement and biases. There is, however, significant evidence that proves the opposite, with these methods leading to harmful discrimination and potentially causing long-lasting negative impacts on customers.

At Lloyds Banking Group, our mission is to be the best bank for customers by combining human and machine intelligence safely and at scale. Our ambition is to work with partners across academia and industry to develop a “Responsible AI” initiative; combining the overlapping fields of Data Ethics, Algorithmic fairness, model explainability and more.

In this talk, we’ll focus on our research on a key issue in credit scoring –avoiding unwanted bias against protected classes. We first explain how discrimination can occur within the different stages involved in supervised machine learning. We then look at the different algorithmic interventions that have been previously proposed by academics, in order to identify whether certain classes within protected groups were treated unfairly. We revise the limitations of each intervention and evaluate them with an example of credit model.

The key benefit of this approach is that it does not involve modifying existing modelling methods. It can be used as a measure of fairness that helps stakeholders take action if a specific class is deemed to have been treated unfairly.