Pioneering new approaches to verifying the fairness of AI models

By supporting, or even automating, decision-making processes, machine learning algorithms can save companies time and money. But they also come with a risk: biases in these AI models can lead to discriminatory decisions. For example, commercially-used facial recognition systems have been shown to work **less accurately on women of colour,** while automated recruitment systems can perpetuate existing prejudices, such as a **negative bias against hiring women**.

For companies to gain the trust of their employees, clients and the wider public, it is crucial for them to be able to prove that their AI models have been developed in as fair a way as possible. A team of researchers led by the Turing's **Ali Shahin Shamsabadi** has now developed a unique framework that can provide companies with **a certificate that verifies the fairness of their model**.

The framework – essentially a pair of algorithms – is applied at the model's training stage, i.e. when the model 'learns' from training data. This avoids potential issues associated with auditing models that have already been trained, such as the auditor testing the model using a dataset that is unrepresentative of the training data. The framework is also completely confidential: the company does not need to reveal its model or training data to the auditor, thus protecting its intellectual property. This is achieved through a cryptographic technique known as 'zero-knowledge proofs', which allow a party to prove statements about its data without revealing the data.

When tested on a type of AI model called 'decision trees' (widely used in sensitive domains such as healthcare and finance), the method could certify fairness in less than two minutes. The researchers are now developing similar certification for two other key aspects of AI decision-making: training data collection and model deployment.

Read the paper: **Confidential-PROFITT: Confidential PROof of FaIr Training of Trees**

> "This is all about trust. Our framework will allow companies to obtain a certificate that shows they are using AI in a fair and ethical way."

Ali Shahin Shamsabadi
Research Associate,
The Alan Turing Institute

Related programmes and teams
Artificial intelligence