Reg. No: 2022/47/D/ST6/01770; Principal Investigator: dr hab. inż. Dariusz Brzeziński

Algorithms and measures for fair and explainable decision systems

In today's highly computerized world, more and more decisions are being made with the help of computers. These tasks are carried out with the help of *decision systems*—algorithms that, taking into account the available data, provide the best decision in a given situation. Decision systems are used in fields such as marketing (in deciding which offers to show to customers), credit risk management (in deciding whether to approve a loan), or insurance (in deciding whether an application should be checked for potential fraud). Fully automated decision-making systems use algorithms from the field of machine learning, where the main task is to discover knowledge from data and later use this knowledge to make a decision. In systems where the final decision is always made by a human, methods from the multiple-criteria decision aid are widely used, where the aim is to show the user a ranking of the best options based on predetermined criteria.

As the volume of available data increases, the aforementioned decision systems are not only becoming more powerful but also more complex. This raises the problem of fully understanding and controlling decision systems in order to maintain fairness and comply with government regulations. These issues were at the heart of the recently proposed EU regulation on AI and the recently adopted UNESCO recommendation on ethics in artificial intelligence. Problems with controlling decision systems also constitute research challenges for scientists, particularly related to the *fairness* and *explainability* of algorithms.

Fairness in artificial intelligence means that the predictive algorithm is not configured to evaluate decisions based on gender, race, or other exclusionary factors. The aim is to ensure that the algorithm or artificial intelligence model is not biased and treats all users equally, without discrimination. However, there are many challenges in achieving fairness in artificial intelligence, including the risk of unexpected behavior of algorithms in the presence of data imbalance, that is, in situations when we have significantly fewer examples of one decision than the other decisions. Another difficulty is defining and measuring fairness in a consistent and objective way.

Explainability is the ability of algorithms to explain their decisions and predictions in a way that is understandable to humans. Explainability helps build trust in decision systems and ensures that they are used in an informed manner. However, achieving explainability can be a challenge because many artificial intelligence algorithms, such as neural networks, are highly complex and difficult to interpret. In addition, explanation methods are most often used for machine learning models and are often not available for decision-support systems, which can also be complex and difficult to interpret.

The proposed project focuses on the two research problems discussed above—explaining decision systems and measuring fairness. As part of the project, we will study the theoretical properties of fairness measures and design new measures that take into account data imbalance. We will also develop methods for visualizing multiple-criteria decision systems that will make it easier for humans to understand the decisions they make. The results of our research will be applied to real-world use cases and made available to the wider public in the form of explainable multiple-criteria decision dashboards and libraries for programmers.