Explainable Artificial Intelligence **Paderborn University**
Summer Semester 2025 Dr. Stefan Heindorf

Exercise Sheet 5

Due date June 23, 2025

1 Task: Case Study

Imagine you are developing a system to predict whether a person should be hired. The data is obtained by extracting data from CVs and stored as a table. The data includes information about age, gender, highest degree, country of citizenship, name, years of experience, skills, and other features. To train and test your system, you have some real-world data on hiring decisions and how the employees performed in the following years, e.g., whether they got promoted, fired, or left the company after a short time.

a. You notice that your hiring system does not perform as well as expected. In production, its performance seems to be lower than on your test set. How would you debug your system? What explanation method would you use? Why?

b. You notice that female applicants receive a lower "hiring score" on average than male applicants. What can you conclude from this? How can you further investigate the fairness of your system? What explanation method would you use? Why?

c. Imagine you applied for a position and received the response from such a hiring system. What should the explanation look like? What explanation method would you use?

d. Imagine you work in the human resource department and you notice that the system makes better predictions than recruiters and managers—not in every case but on average. You would like to gain insights into how this system works and to improve your hiring process. What explanation method would you use? Why?

2 Task: Neural Network Explanations

Experiment with the Captum library[1] and try to execute the tutorial "Model Interpretation for Pretrained ResNet Model".[2] The required image of a swan can be downloaded from `https://github.com/pytorch/captum/tree/master/tutorials/img/resnet`.

[1] `https://captum.ai/`
[2] `https://captum.ai/tutorials/Resnet_TorchVision_Interpret`