Artificial Intelligence for Sciences and Humanities

Homework 8: Bias in AI Systems

DO NOT POLLUTE! AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

In this homework you will explore the vulnerabilities of AI systems by introducing a harmful bias on the neural network you implemented in Homework 7, where you loaded the MNIST dataset using the following code:

```
mnist = fetch_openml('mnist_784', version=1, as_frame=False)

X = mnist.data
X = X.astype(np.float32) / 255.0
y = mnist.target.astype(int)

print("X shape:", X.shape)
print("y shape:", y.shape)
```

This code loads the labels in the vector **y**. We are going to introduce a harmful bias to our AI system in a very simple way: we will systematically alter the labels of some samples. Each problem is worth 40 points.

Problem 1. Load the MNIST dataest using the code above, and then write code to replace *all* labels in **y** that are equal to 7 with a 1. Keep track of the indices (samples) that you changed, as you will need this later. Deliver your code.

Problem 2. Repeat the necessary steps of Homework 7 to train your network using the altered vector **y**.

Problem 3. Use your AI system on a few test (unseen) samples that were initially 7's. Deliver your code and sample images with their corresponding outputs.

Problem 4. Did your system perform satisfactory? Explain in your own words what happened.

Problem 5. Describe a scenario where this type of bias can be used unethically in a modern AI system like ChatGPT.