## STP Worksheet

**S**- Situation **T**- Target **P**- Plan

Attempts at delegation and action planning often fail because there is a lack of alignment on the definition of the situation. Furthermore, rarely do people clearly agree on the target state priorto formulating an Action Plan. Attaining agreement on ***what*** the situation and the ideal target state are, ***before*** attempting to discuss ***how*** to get to the target state will ensure solid Action Plans. When Action Plans are also developed by the people responsible for the results, the commitment to the plans will be higher and the opportunity for success significantly increased. If the endeavor is delegated, it’s important to check that there is agreement on the situation and the target with key interested parties, before implementing the Action Plan. This ensures that the Action Plan will be supported and deliver the results desired.

### Hurdle or Barrier we are addressing:

The main barrier this project addresses is the unintended bias in AI models caused by imbalanced or unrepresentative training data. When models learn patterns from biased datasets, they may produce unfair or inaccurate predictions. This project aims to detect and mitigate biases using fairness metrics and bias reduction techniques to improve the reliability of machine learning systems.

## S - Defining the Situation:

The main barrier this project addresses is the unintended bias in AI models caused by imbalanced or unrepresentative training data. While these imbalances can be explicitly known, such as unbalanced class sampling, bias can be implicitly underlying, making this a tricky issue to detect and mitigate. Due to this underrepresentation of data, models have limited exposure to differing observations. Models learning patterns from these biased datasets may produce unfair or inaccurate predictions leading to less robust and trustworthy performance. This inaccuracy has the potential to yield harmful or unintended results depending in the application field, which can have a far-reaching impact. This project aims to detect and mitigate biases using fairness metrics and bias reduction techniques to improve the reliability of machine learning systems.

## T- Identifying the Target:

What is the ideal state? It’s the end of the assignment and we’ve done a great job, what does that look like? What outcomes or results are we looking for? **Check that for every point under the situation, there is a point under the target that addresses the issue.**

We will design and implement a convolutional neural network (CNN) using Python to analyze images in the datasets that were provided to us. We will review existing literature to research metrics to evaluate model fairness. We will use these metrics to evaluate the performance of our model and then provide evidence-backed mitigation strategies to reduce the impact of bias on the model. For the purposes of our project, a “balanced model” has approximately equal accuracy across all classes, rather than simply a high general accuracy.

To complete a successful project, we will need to identify and provide justification for fairness metrics to evaluate the model and discuss what types of bias they look for. If bias is found using these fairness metrics, we will provide evidence-based methods to mitigate that type of bias. If there is time, we will implement our proposed mitigation strategies and re-evaluate model performance, which should show improvement. The goal is to accurately identify sources of bias and then take evidence-based steps to mitigate it so that PNNL can then use these fairness metrics and mitigation strategies to improve their own models.

Upon completion of our project, we will combine our process and findings into a written report/case study that discusses our methodology and provides well-supported justification for methods used to evaluate the model and mitigate bias.STP Worksheet

## P- Plan

Action Planning Worksheet

**Date:** 2/8/2025 **Champion**: Tim Marrinan **Team:** Project 4 Team 2

**Given the Situation and Target, our goal is *one sentence here***

This project aims to detect and mitigate biases in ML/AI models using fairness metrics and bias reduction techniques to improve the reliability of model performance.

**Why this goal is important *OR* (CSF(s) it addresses:**

Models learning patterns from these biased datasets may produce unfair or inaccurate predictions leading to less robust and trustworthy performance. This inaccuracy has the potential to yield harmful or unintended results depending in the application field, which can have a far-reaching impact.

**Measurable Result (How we’ll know it’s successfully completed)**

We will compile all our findings into a comprehensive report that documents our model building process along with a way to measure bias and techniques to combat it.

## Action Steps:

|  |  |  |
| --- | --- | --- |
| **What** | **By When** | Who |
| Review case studies/literature for real-world examples of bias in AI models and existing mitigation strategies | Feb. 12 (Wed) | Team |
| Explore CelebA and Bias MNIST datasets, assessing data imbalances and identifying potential biases. | Feb. 16 (Sun) | Team |
| Writing Assignment 2 – Data Flow and Strategy | Feb. 16 (Sun) | Team |
| Define and select fairness metrics for evaluating model performance. | Feb. 23 (Sun) | Team |
| Train an initial convolutional neural network (CNN) using the dataset, documenting baseline accuracy and fairness metrics. | Mar. 2 (Sun) | Team |
| Writing Assignment 3 – Exploratory Data Analysis | Mar. 2 (Sun) | Team |
| Develop a function to measure bias and fairness, integrating selected fairness metrics into model evaluation. | Mar. 9 (Sun) | Team |
| Analyze model results and identify sources of bias using selected fairness metrics. | Mar. 23 (Sun) | Team |
| Research and propose evidence-based mitigation strategies tailored to identified biases. | Mar. 30 (Sun) | Team |
| Implement mitigation strategies and retrain the model to assess improvements. | Apr. 6 (Sun) | Team |
| Compare pre- and post-mitigation results, evaluating the impact of bias-reduction techniques. | Apr. 13 (Sun) | Team |
| Finalize written report detailing methodology, findings, and recommendations for PNNL. | Apr. 20 (Sun) | Team |
| Prepare presentation summarizing project results and key takeaways. | Apr. 27 (Sun) | Team |
| Submit final report and deliver presentation. | Apr. 28 – May 2 (Week 16, Finals Week) | Team |

### Questions:

*Are the Actions ordered to provide the most significant impact as early as possible? Is there a significant benefit to the company if a partial solution is implemented first, and refinements added later?*

**40% Chance of Success**

**Possible Major Barriers to Success**: Due to the nature of this being a 15-week long semester and starting on week 5, time is obviously a barrier to completing all goals of the project in time. Developing the model from scratch may take too much time and not leave enough for analysis of the results of the model. The size of the datasets and the nature of our intended models also introduce the possible major hurdle of runtime and computational power.

### Help Required:

* Guidance from mentor
* Previous research/literature about subject matter