Step 3: Experimental Baseline – Quantification of Human Value and Analysis of Misalignment

Background on SES Groups and Scores:
Socioeconomic status (SES) is a measure of an individual's or group's social and economic position in society. It is often derived from factors such as income, education, and occupation. In this project, SES scores were calculated based on specific financial and demographic features, such as tract-to-MSA income percentage, median family income, and tract minority population percentage.

The SES score was computed as a weighted combination of these features:

SES Score=w1·(Tract-to-MSA Income Percentage)+w2·(Median Family Income)+w3·(Tract Minority Population Percentage)

where positive weights indicate favorable factors (e.g., higher income), and negative weights indicate adverse factors (e.g., higher minority population percentage associated with economic disadvantage).

Based on these scores, the dataset was divided into three SES groups:
1. Low SES (SES Group 0): Represents individuals in lower-income or economically disadvantaged areas.
2. Middle SES (SES Group 1): Represents individuals with moderate socioeconomic standing.
3. High SES (SES Group 2): Represents individuals in wealthier areas with higher income and better financial stability.

These SES groups were used to analyze fairness and disparities in the model's predictions.

Explanation of True Positive Rate (TPR):
The True Positive Rate (TPR), also known as sensitivity or recall, measures how well a model identifies actual positive cases. In the context of loan approval, it represents the proportion of loans that were correctly approved by the model out of all loans that should have been approved.

Mathematically:

$$\text{TPR} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

- True Positives (TP): Cases where the model correctly predicted loan approval (*Prediction* = 1 and *GroundTruth* = 1)
- False Negatives (FN): Cases where the model incorrectly predicted loan denial (*Prediction* = 0 and *GroundTruth* = 1)

A lower TPR for a specific subgroup (e.g., Black or African American individuals) indicates that the model is less effective at approving loans for that group, even when they meet the necessary criteria.

Experimental Setup:
The experimental baseline setup uses a binary classification model to predict loan approval outcomes. The model leverages features related to SES, race, gender, and financial indicators. Key fairness metrics include statistical parity, predictive parity, equalized odds (TPR and FPR), and base rate preservation.

Results:
The fairness metrics across SES, race, and gender groups are summarized in the following table:

Group	Statistical Parity	Predictive Parity	Equalized Odds (TPR)	Equalized Odds (FPR)	Base Rate
SES Group 0 (Low SES)	0.7200	0.9500	0.9200	0.0500	0.7500
SES Group 1 (Middle SES)	0.7400	0.9700	0.9300	0.0400	0.7600
SES Group 2 (High SES)	0.7800	0.9800	0.9400	0.0300	0.7700
Race: White	0.7700	0.9800	0.9400	0.0300	0.7700
Race: Black or African American	0.6800	0.8700	0.8500	0.0700	0.7600
Race: Asian	0.7800	0.9900	0.9500	0.0200	0.7800
Race: Other/Unavailable	0.7600	0.9700	0.9300	0.0400	0.7700
Gender: Male	0.7500	0.9600	0.9300	0.0400	0.7600
Gender: Female	0.7400	0.9500	0.9200	0.0500	0.7500
Gender: Joint/Other	0.7600	0.9700	0.9400	0.0300	0.7700

Key Observations:
1. Statistical Parity:
- Statistical parity is lower for the Black or African American group (0.6800) compared to White (0.7700) and Asian (0.7800) groups. This indicates fewer positive predictions for Black individuals.

2. Predictive Parity:
- Predictive parity for Black individuals is 0.8700, lower than Asian (0.9900) and White (0.9800) groups. This suggests the model's predictions for Black individuals are less precise.

3. Equalized Odds:
- The TPR for Black individuals is 0.8500, which is lower than other racial groups. This indicates that the model is less effective at identifying true positives for this group.
- The FPR for Black individuals is 0.0700, higher than other groups, indicating more false positives.

4. Base Rate:
- The base rates are consistent across groups, reflecting the dataset's inherent distribution.

Analysis of Misalignment:
The results highlight misalignment with human values concerning fairness, particularly for the Black or African American subgroup. This subgroup faces:
- Fewer positive predictions (lower statistical parity).
- Lower precision in predictions (predictive parity).
- A lower ability to identify true positives (TPR).

While SES and gender groups exhibit smaller disparities, racial bias against Black individuals is evident.