AI Tools Assignment — Part 3: Ethics & Optimization

Name: Njambi Hinga
Course: AI for Software Engineering — Power Learn Project
Date: October 2025

A. Reflection on AI Bias and Fairness

Bias in AI systems often originates from **imbalanced or unrepresentative datasets**, where certain classes or groups dominate the training data.
In my Scikit-learn Iris classifier, for example, the dataset was balanced, but in real-world datasets, skewed samples can lead to **overfitting** toward majority classes.

For the MNIST CNN model, bias may occur if certain handwriting styles are underrepresented, causing poor generalization.
In NLP tasks with spaCy, sentiment models may inherit **linguistic bias** if the training corpus overrepresents specific demographics or writing tones.

Mitigation strategies include:

- Data rebalancing using **SMOTE** or class weights
- Using **Fairness Indicators** or **IBM AI Fairness 360** to evaluate model bias
- Applying **LIME** or **SHAP** for interpretability and transparency

Responsible AI means continuously auditing both the data and model decisions to ensure **fairness, accountability, and transparency**.

B. Debugging and Optimization Task

Below is an example of a faulty TensorFlow model.
The model fails to train because it lacks activation functions and uses an incorrect loss function.

In [1]:
```
import tensorflow as tf

# Buggy Model
model = tf.keras.Sequential([
    tf.keras.layers.Dense(10, input_shape=(784,)),
    tf.keras.layers.Dense(10)
])
```

```
model.compile(optimizer='sgd', loss='mse', metrics=['accuracy'])
model.fit(x_train, y_train, epochs=5)
```

```
c:\Users\NATASHA\OneDrive\Desktop\AI_Tools_Assignment\venv\Lib\site-packages\keras\src\layers\core\dense.py:92: UserWarning: Do not pass an `input_shape`/`input_dim` argument to a layer. When using Sequential models, prefer using an `Input(shape)` object as the first layer in the model instead.
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
---------------------------------------------------------------------------
NameError                                 Traceback (most recent call last)
Cell In[1], line 10
      4 model = tf.keras.Sequential([
      5     tf.keras.layers.Dense(10, input_shape=(784,)),
      6     tf.keras.layers.Dense(10)
      7 ])
      9 model.compile(optimizer='sgd', loss='mse', metrics=['accuracy'])
---> 10 model.fit(x_train, y_train, epochs=5)

NameError: name 'x_train' is not defined
```

🧩 Problems:

1. Missing activation functions → model can't learn non-linear features
2. Wrong loss function → `mse` is not suitable for classification
3. Output layer missing `softmax` activation

✅ Fixed Code:

In [2]:
```
model = tf.keras.Sequential([
    tf.keras.layers.Dense(128, activation='relu', input_shape=(784,)),
    tf.keras.layers.Dense(64, activation='relu'),
    tf.keras.layers.Dense(10, activation='softmax')
])

model.compile(optimizer='adam',
              loss='sparse_categorical_crossentropy',
              metrics=['accuracy'])

# model.fit(x_train, y_train, epochs=5, batch_size=32, validation_split=0.2)
print("✅ Model compiled successfully!")
```

```
✅ Model compiled successfully!
```

C. Fairness and Optimization Techniques

Category	Example Techniques
Model Optimization	Learning rate tuning, early stopping, dropout regularization
Performance Monitoring	TensorBoard, Keras callbacks
Fairness Tools	TensorFlow Fairness Indicators, IBM AI Fairness 360

Category	Example Techniques
Explainability	LIME, SHAP
Data Quality	Remove duplicates, normalize, augment minority classes

Ethical optimization ensures that AI systems perform well **without compromising fairness or transparency**.

🏁 Conclusion

Through this exercise, I explored ethical considerations, fairness, and optimization strategies for AI systems.
Building responsible AI models requires balancing accuracy with fairness and explainability.