**Ethical considerations**

**Identify potential biases in your MNIST or Amazon Reviews model. How could tools like TensorFlow Fairness Indicators or spaCy’s rule-based systems mitigate these biases**

**MNIST (hand‑written digits)**

**Bias**

Sampling bias – writers are mainly US high‑school students from the late‑1990s; little variation in age, culture, script style.   
Class imbalance – “1” and “0” are over‑represented; “5” and “9” under‑represented.   
Image quality bias – darker strokes are easier for CNN filters to detect.

**How Tensorflow mitigate bias**

Slice the evaluation set by writer ID, digit class, or stroke darkness quartile and compare false‑negative and false‑positive rates per slice. Fairness Indicators will surface any slice where the error gap exceeds a τ‑threshold you choose and plot it in an interactive dashboard

**How spacy’s rule based system mitigate bias**

• Before training, use spaCy’s rule matcher to tag digits whose bounding‑box ink entropy < τ (light strokes) and up‑sample them.  
• During post‑processing, add a simple rule: if the logit gap between top‑2 classes is < 0.05 and the digit contains a closed loop, favour {0,6,8,9

**Amazon review models**

Bias

Genre bias – electronics vs. romance novels use different sentiment cues.  
Demographic and dialect bias – AAVE or reviews with gender‑coded words (e.g. “bossy”, “gentle”).  
Star rating and length bias – very short 1‑star rants vs. long positive reviews.

**How Tensorflow mitigate bias**

Slice by **product category**, **review length buckets**, **dialect detection flag**, **gender‑coded vocabulary presence**. Fairness Indicators computes metrics such as Positive Predictive Value and Equal Opportunity per slice and lets you drill down to inspect mis‑classified examples, then export the slices that need re‑balancing

**How spacy’s rule based system mitigate bia**

 Create pattern lists (e.g. r‑AAVE patterns, gendered adjectives). spaCy’s Matcher/PhraseMatcher can tag these spans on‑the‑fly, letting you: 1) augment or re‑weight under‑represented slices during training; 2) override model output when a rule triggers; 3) log flagged cases for human review.