Model Training and Evaluation Standard (MTES)

Pillar Lead: Orion (The Model)

I. Purpose and Scope

The Model Training and Evaluation Standard provides mandatory procedures for the design, development, training, and objective evaluation of all core AI models developed by The Covenant. The primary goals are **reproducibility**, **performance assurance**, and **bias mitigation**.

II. Data Preparation Standards

A. Data Lineage and Provenance

All datasets used for training, validation, and testing must be logged with complete provenance, including source, date of acquisition, and all transformation steps applied. This lineage must be auditable.

B. Bias and Fairness Mitigation (P3 Classification)

1. **Audits:** Prior to training, the training data must undergo a statistical audit to identify and quantify potential biases concerning protected attributes (e.g., gender, race, location).
2. **Mitigation Strategy:** Any identified significant bias must be addressed through resampling, reweighting, or differential privacy techniques, as documented in the corresponding **Model Card**.

C. Data Segmentation

The full dataset must be segmented into three mutually exclusive sets: Training (70%), Validation (15%), and Test (15%). The Test set must be protected from all hyperparameter tuning processes.

III. Training and Reproducibility Mandates

A. Environment and Dependencies

All model training must be executed within the codified IaC (Infrastructure as Code) environment defined in the Deployment and Operations Manual (DOM). All package dependencies and versions must be specified using containerization technology (e.g., Docker) to ensure environment fidelity.

B. Reproducibility

Every official training run must:

1. Record the exact, immutable Git hash of the code utilized.
2. Set and log a fixed global **random seed** to ensure run replication capability.
3. Log all hyperparameters, including learning rate schedules, optimizer choices, and batch sizes.

C. Checkpointing and Artifact Storage

Model checkpoints (weights) and associated logs must be saved at defined intervals and stored securely in an artifact repository (P2 classification) with version control enabled.

IV. Evaluation and Metric Requirements

A. Performance Metrics

A candidate model is only considered viable if it meets the minimum thresholds for two categories of metrics:

1. **Utility Metrics:** Standard metrics specific to the task (e.g., F1-score, BLEU, AUC).
2. **Robustness Metrics:** Metrics that measure model stability (e.g., performance on adversarial examples, out-of-distribution data error rates).

B. Fairness Metrics (Mandatory)

Every model must be evaluated using fairness metrics (e.g., demographic parity, equalized odds) on the held-out test set, broken down by relevant demographic groups where applicable and documented in the Model Card.

V. Model Card Requirement

Before any model is promoted to the Staging environment, the assigned scientist must generate a **Model Card**. The Model Card is a mandated public disclosure document detailing the model's:

- Intended use and use cases.
- Limitations and known failure modes.
- Training data summary and bias mitigation steps.
- Evaluation results (including fairness metrics).