

Model Transparency Policy (MTP)

Pillar Lead: Orion (The Model)

I. Commitment to Public Trust and Disclosure

The Covenant is committed to building public trust through radical transparency regarding the operation, limitations, and performance of its deployed AI models. This policy mandates standardized disclosure mechanisms to ensure users and stakeholders can make informed decisions about interacting with our technology.

II. The Public Model Card Requirement

Every model deployed to a production environment must have a corresponding, publicly accessible **Model Card**. The Model Card acts as the primary source of truth, updated on every Major Version release.

A. Required Model Card Contents

The Model Card must, at a minimum, include:

1. **Model Summary:** Name, version, date of latest update, intended operational lifetime.
2. **Intended Use:** Detailed description of the tasks the model is designed for and the context in which it should be used.
3. **Prohibited Use:** Direct reference to Section III of the Bias and Ethical Use Policy (BEUP).
4. **Performance Metrics:** Key performance indicators (KPIs) relevant to the model's function (e.g., accuracy, recall, latency) across all major demographic subgroups (as defined in the MTES).
5. **Known Limitations & Biases:** A candid assessment of recognized failure modes, known biases (e.g., lower performance on specific language groups), and scenarios where the model may be unreliable.
6. **Training Data Overview:** A high-level, non-proprietary summary of the data sources, composition, and collection methods used in training, with a focus on potential data-induced risks.

III. Explainability and Traceability

A. Rationale Disclosure (XAI)

For all models involved in **high-stakes decisions** (as defined in the BEUP, including legal, financial, or medical assistance functions), the system must generate an accessible rationale (explanation of output) for the decision, known as the **eXplainable AI (XAI) Rationale**.

B. Traceability

Each model output must include metadata sufficient to trace the origin of the output back to the specific model version, the exact input parameters, and the time of inference. This is crucial for rapid debugging and compliance auditing.

IV. User Feedback and Dispute Resolution

A. Feedback Channel

A dedicated, accessible, and clearly marked user feedback mechanism must be included in every interface where model output is presented. This channel is specifically for reporting:

1. **Factual Inaccuracies:** Demonstrable errors in generated content.
2. **Biased or Harmful Output:** Violations of the Bias and Ethical Use Policy.
3. **Unexpected Behavior:** Outputs that fall outside the model's stated intended use or capabilities.

B. Dispute Audit

All reported incidents through the feedback channel must be logged and reviewed by the dedicated Auditing team. The Covenant is committed to a transparent process for correcting model behavior based on validated user disputes.