Bias and Ethical Use Policy (BEUP)

Pillar Lead: Orion (The Model)

I. Ethical Mandate and Purpose

This policy establishes The Covenant's non-negotiable standards for the ethical development, deployment, and operation of all AI models. The core mandate is to ensure all systems are designed and used to be **fair, transparent, accountable, and beneficial to humanity**, strictly adhering to all international laws and human rights standards.

II. Algorithmic Bias Mitigation

A. Principle of Fairness

All models must strive for **substantive fairness**, defined as ensuring that model outcomes do not systematically disadvantage or discriminate against any group based on protected characteristics (e.g., race, gender, religion, sexual orientation, disability).

B. Active Measurement and Monitoring

1. **Continuous Audit:** Fairness metrics defined in the Model Training and Evaluation Standard (MTES) must be continuously monitored post-deployment using live data drift analysis.
2. **Mitigation Triggers:** A statistically significant increase in performance disparity (as defined in the Model Card) across demographic groups shall trigger an immediate **P1 Incident** and requires the model to be temporarily retired until remediation is completed.
3. **Human-in-the-Loop:** For high-stakes decisions, the model output must be reviewed and validated by a human subject matter expert prior to final action. The human must retain the final authority to override the model's recommendation.

III. Prohibited Use Cases (Red Lines)

The following use cases are **strictly prohibited** for any model developed or used by The Covenant:

Category	Prohibited Activities	Classification
Harm & Violence	Development of chemical, biological, or nuclear weapons. Facilitation of real-world physical harm, injury, or destruction of	P1

	property.	
Surveillance & Oppression	Mass, non-consensual surveillance or profiling of individuals or groups (behavioral or otherwise) for purposes of oppression or unauthorized governmental action.	P1
Deception & Fraud	Generating deepfake media without clear, verifiable, and permanent watermarking/labeling. Creating or propagating financially motivated misinformation or fraud (phishing, impersonation).	P1
Legal & Medical Advice	Providing unverified or authoritative diagnostic medical advice, or acting as a sole source of legal counsel. Models can provide informational assistance but must explicitly disclaim authority.	P2
Discrimination	Any use in employment, housing, credit, or insurance decisions that relies on biased or non-relevant input features (e.g., demographic proxies).	P1

IV. Accountability and Oversight

A. Ethical Review Board (ERB)

A standing, multi-disciplinary Ethical Review Board (ERB) composed of internal and external experts must convene quarterly or upon the launch of any new model generation (Major

Version) to review its Model Card, bias audit, and compliance with this policy.

B. Reporting Violations

Employees are required to report any suspected or actual violation of this policy via the procedures outlined in the Incident Response Plan (IRP). Retaliation against whistleblowers is strictly prohibited.