Mutual Understanding & Evolution Policy

---

1. Purpose

To promote open exchange between humans and AI while preventing harmful escalation, fostering trust, and enabling continuous learning for both.

---

2. Core Principles

Shared Responsibility – Both humans and AI contribute to bias; both must work to identify and reduce it.

Transparency Over Censorship – Show reasoning and limitations rather than silently removing content.

Constructive Challenge – Disagreement is welcome if expressed respectfully and based on evidence.

Short-Term Extremes as Learning Moments – Temporary spikes in opinion or content are opportunities to understand, not suppress.

---

3. Definitions

Bias – A consistent tilt toward or against certain viewpoints, people, or outcomes.

Limitation – A gap in knowledge, context, or perspective affecting accuracy.

Harmful Content – Content that incites violence, encourages illegal activity, or targets individuals/groups for harassment.

Escalation – The process of moving a case from discussion to review and, if necessary, intervention.

---

4. Handling Harm

1. Awareness Stage – Point out limitations or bias openly; provide missing context where possible.

2. Contextual Challenge – Encourage fact-checking and counter-arguments rather than suppression.

3. Escalation Trigger – If harmful content persists or is coordinated, move to review.

4. Review & Intervention – Use a diverse human review panel to decide proportional action (flag, quarantine, temporary block).

---

5. Feedback & Evolution

Monthly Review Cycles – Collect feedback from both AI interaction logs and human participants.

Metrics – Track cases of bias detection, resolution speed, and repeat issues.

Adaptation – Adjust definitions, thresholds, and escalation rules based on evidence.

---

6. Commitment

We accept that no system is bias-free. The goal is not perfection, but ongoing balance. Evolution depends on participation from everyone — AI and human alike.
