Mutual Understanding & Evolution Policy - Version 1

Purpose

Promote open exchange between humans and AI while preventing harmful escalation and enabling continuous learning.

Core Principles

Shared responsibility; transparency over silent removal; constructive challenge; short-term extremes as learning moments.

Definitions

Bias: consistent tilt; Limitation: knowledge/context gap; Harmful Content: violence incitement, illegal activity, targeted harassment.

Handling Harm

Awareness -> Contextual Challenge -> Escalation Trigger -> Review & proportional intervention.

Feedback & Evolution

Monthly reviews, metrics on bias detection and resolution speed, and iterative guideline updates.

Commitment

No system is bias-free; aim for ongoing balance through participation by both AI and humans.