Aegis-R: Threat Model & Failure Modes

Threat Model

Adversaries may attempt to poison learning inputs or manipulate telemetry.

Attackers may attempt slow behavioral drift to normalize malicious actions.

Mitigations

Separation of reasoning and learning planes.

Immutable attack logic and signed governance rules.

Failure Modes

Total telemetry compromise reduces effectiveness.

Incorrect human approvals may promote unsafe assumptions.

Design Tradeoffs

Aegis-R prioritises correctness over speed.

Human involvement is mandatory for trust promotion.