§6 Effective Oversight

Transparency and controllability

Trace and Explain | Intervene and Correct | Audit and Govern

Transparency · Interpretability · Controllability · Governance

§4 Behavioral Constraints

Behavioral constraints, alignment, and fairness

- Behavioral Safety
- Value Alignment
- Group Fairness

Primarily endogenous risks

§2 Recursive Improvement

Models & agent systems

Experience & Feedback: E_{t+1} F_{t+1}

Editable Components: Weights · Prompts · Memory · Skills / Code

S1 Diagnose → S2 Propose → S3 Evaluate → S4 Implement → S1 Diagnose

Improvement cycle

RI: Apply S1–S4 to the bias update function

$$\mathcal{U}_t \rightarrow \mathcal{U}_{t+1}$$

$(h_t, B_t, \mathcal{U}_t)$ → $(h_{t+1}, B_{t+1}, \mathcal{U}_{t+1})$ → ...

Accepted revisions persist into subsequent cycles

§5 System Security

Resilience, privacy and data governance

- Security Resilience
- Privacy Protection
- Secure Inheritance

Primarily exogenous risks

§3 Improvement Validity

Genuine and sustained improvement

Measure and Verify | Retain and Transfer | Compare and Isolate

Validity · Reliability · Robustness · Truthfulness

§7 Evaluation and Evidence

Individual update | Successor Inheritance | Mechanism Revision

§8 Applications

- Safety Harness Engineering
- Coding & Retrieval
- Healthcare & Robotics

§8 Open Problems

- Persistent Constraints
- Revised Improvers Assurance
- Long-Horizon Assurance