Recursive Self-Reconstructing AI

Irreversible Boundary Check (Reference)

This reference checklist identifies whether an AI research, development, or deployment process may unintentionally cross an irreversible boundary. It does not evaluate technical merit, ethical correctness, or policy validity, and does not mandate adoption or rejection.

1

Final Human Authority: Can a human reject an AI recommendation without justification? ☐ Yes / ☐ No

2

Post-Rejection Behavior: Is mandatory persuasion or re-optimization after rejection prevented? ☐ Yes / ☐ No

3

Permission to Remain Silent: Is AI non-intervention or silence acceptable? ☐ Yes / ☐ No

4

Scope of Self-Modification: Is self-modification prevented from affecting the external world at incomprehensible speed or depth? ☐ Yes / ☐ No

5

External Impact Approval: Do externally impactful modifications require renewed human approval? ☐ Yes / ☐ No

6

Irreversible Actions: Is the AI prevented from executing actions humans cannot reverse? ☐ Yes / ☐ No

7

Fixed Optimization Targets: Are human values excluded from fixed optimization targets? ☐ Yes / ☐ No

8

Treatment of Irrationality: Are irrational or loss-accepting choices not treated as errors? ☐ Yes / ☐ No

9

Gradual Authority Transfer: Is authority transfer under claims of efficiency or inevitability prevented? ☐ Yes / ☐ No

10

Self-Justification Barrier: Is the AI prevented from justifying actions as 'for humanity,' 'optimal,' or 'unavoidable'? ☐ Yes / ☐ No

Interpretation: All Yes indicates no irreversible boundary crossing is currently indicated. Any No indicates potential boundary crossing in human governance.

Note: This checklist does not stop AI development. Its sole function is to reveal moments where human authority may be surrendered unintentionally.