

Toward Trustworthy RSI

Ensuring AI systems remain trustworthy as they recursively improve themselves

Toward
Trustworthy RSI

Transparency &
Controllability

Effective Oversight

Behavioral Constraints,
Alignment & Fairness

Constraint Preservation

Security, Privacy
& Data Governance

Constraint Preservation

Genuine & Sustained
Improvement

Improvement Validity

Recursive
Self-Improvement

System and updater changes
can alter trustworthiness

Evaluation & Evidence

Misleading
Gains

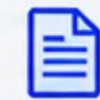
Unsafe & Unfair
Outcomes

Inherited
Compromise

Privacy
Leakage

Weakened
Oversight

Trustworthiness Claims



Update Level

Is the accepted update trustworthy?



Successor Level

Do gains and requirements persist?



Meta-Improvement Level

Does the revised updater produce better, trustworthy updates?

Human
Oversight