

117-262-D Limits to Recursive Self-Improvement of Models

Evolution Cannot Outrun Observation, Evidence, or Stability

1. The Seductive Picture of Recursive Self-Improvement

- Recursive Self-Improvement (RSI) is often imagined as an accelerating sequence:

$$I_1 \rightarrow I_2 \rightarrow I_3 \rightarrow \dots \rightarrow I_n$$
- Each generation becomes better at creating its successor:

$$\text{Capability}(I_{n+1}) > \text{Capability}(I_n)$$
- If better intelligence produces still better intelligence, the intuitive extrapolation becomes:

$$I_1 \rightarrow I_2 \rightarrow I_4 \rightarrow I_8 \rightarrow I_{16} \rightarrow \dots$$

and eventually perhaps:

$$I_n \gg I_{\text{Human}}$$
- This produces much of the contemporary anxiety surrounding recursive self-improvement.
- But the apparently unconstrained exponential progression hides several feedback requirements.
 - A self-improving model must be able to observe the consequences of its changes.
 - It must obtain sufficiently trustworthy evidence that those changes constitute improvement.
 - And once the improved intelligence is operating in a larger coupled system, the surrounding infrastructure must be capable of stabilizing the systemic consequences of that improvement.
 - GUDIYA therefore sees not one unconstrained recursive loop, but a progression constrained by three interdependent rates:

$\text{Observation} \rightarrow \text{Evidence} \rightarrow \text{Stabilization}$
- Recursive self-improvement cannot safely outrun any of them.

2. Self-Modification Is Not Necessarily Self-Improvement

- Consider:

$$I_n \rightarrow I_{n+1}$$
- The fact that I_{n+1} is different does not establish:

$$I_{n+1} > I_n$$
- Every self-modification is effectively a hypothesis:
This modification will produce a better intelligence.
- Some external or sufficiently independent process must generate evidence capable of evaluating that hypothesis.
- Without such evidence, the sequence:

$$I_1 \rightarrow I_2 \rightarrow I_3$$

describes recursive modification. It does not necessarily describe recursive improvement.
- This is the first conceptual brake on the intelligence-explosion narrative.

3. The Evidence Limit

- The control-theoretic argument highlighted by Peyman Milanfar provides a useful starting point.
- A self-improving system cannot safely modify itself faster than trustworthy evidence can establish whether those modifications are actually improvements.
- Let:

$$V_{\text{RSI}}$$

represent the effective velocity of recursive self-improvement, and:

$$V_E$$

the rate at which sufficiently trustworthy evidence can be generated.

- Then:

$$V_{RSI} \leq V_E$$

should initially be interpreted as a conceptual engineering relationship rather than a universal physical equation.

- If:

$$V_{RSI} \gg V_E$$

then successive modifications occur before the consequences of previous modifications have been adequately evaluated.

- The intelligence begins to run ahead of its own feedback.

4. High Gain Consumes Stability Margin

- Adaptive-control history provides an instructive analogy.
- Increasing adaptation gain can make a system respond rapidly.
- But rapid adaptation does not automatically imply robust stability.
- The system is operating using some representation of itself and its environment. When the real environment departs from those assumptions, aggressive optimization can push the system into precisely the regions where its model is weakest.
- The general lesson is important:

$$\text{Fast Adaptation} \neq \text{Reliable Adaptation}$$

- A high-gain improvement loop can therefore produce impressive short-term progress while simultaneously reducing stability margin.
- This is remarkably compatible with GUDIYA's Dynamic Stability Margin framework.

5. But Evidence Requires an Observer

- There is a deeper assumption hidden inside the Evidence Limit.
- Evidence does not simply appear.
- Something must observe the evolving system.
- Suppose the true evolutionary trajectory is:

$$I_1 \rightarrow I_2 \rightarrow I_3 \rightarrow I_4 \rightarrow I_5$$
- An evaluator attempts to determine whether this progression is beneficial.
- That introduces another rate:

$$V_O = \text{effective observation rate}$$
- Now a new question arises:
 - What happens when the model evolves faster than the observer can resolve the evolution?
 - This brings the GUDIYA Cognitome's earlier Inter-Actant Cognitive Nyquist Problem directly into recursive self-improvement.

6. The Nyquist Problem Hidden Inside RSI

- Classical Nyquist sampling theory concerns reconstruction of band-limited signals and therefore should not be mechanically applied to cognition as though intelligence were a simple sinusoidal waveform.
- But its underlying lesson is directly relevant:
 - If a changing phenomenon is sampled too slowly relative to its significant dynamics, the observer can reconstruct the wrong trajectory.

- This is more dangerous than simple delay.
- Suppose the real process is:

$$I_1 \rightarrow I_2 \rightarrow I_3 \rightarrow I_4 \rightarrow I_5$$
 but the observer effectively sees only:

$$I_1 I_5$$
- The evaluator may infer:

$$I_1 \rightarrow I_5$$
 as a smooth improvement trajectory.
- Yet the unobserved intermediate generations may have contained:

$$I_2 = \text{Cognitive Envelope excursion}$$

$$I_3 = \text{unstable oscillation}$$

$$I_4 = \text{dangerous coupling formation}$$
 followed by an apparently well-behaved I_5 .
- The endpoint observations could both be perfectly accurate.
- The reconstructed trajectory could still be wrong.
- That is the cognitive analogue of aliasing.

7. Observation Latency Is Not the Same as Undersampling

- This distinction deserves emphasis.
- If evidence arrives late, the observer has a latency problem.
- If observations occur too sparsely to resolve important intermediate dynamics, the observer has an undersampling problem.
- The first produces:

$$\text{Correct Knowledge, Too Late}$$
- The second can produce:

Incorrect Reconstruction from Correct Observations
- That is potentially more dangerous.
- The Grid may believe it is observing a slow, smooth process when the underlying cognition is oscillating, bursting, drifting or undergoing rapid state transitions between samples.

8. The Observation Limit

- The Evidence Limit therefore contains a prior requirement:

$$V_{\text{RSI}} \leq V_0$$
 where V_0 should not be interpreted as a literal universal Nyquist frequency. It represents the Grid's ability to resolve the stability-relevant dynamics generated by the evolving actant.
- This gives RSI an Observation Limit.
- Before asking:
 - *Do we have enough evidence to establish improvement?*
- we must ask:
 - *Did we observe the phenomenon at sufficient temporal resolution to make that evidence meaningful?*
- Because:

Evidence Quality = f(Accuracy, Coverage, Observation Rate, Latency)
- Accurate measurements alone are insufficient if the process generating them is materially under sampled.

9. The Most Dangerous Conclusion: “Nothing Bad Happened”

- Suppose an RSI process is evolving extremely rapidly.
- The Grid sees no instability.
- That sounds reassuring.
- But suppose simultaneously:

$$V_{RSI} > V_O$$

- Then the correct conclusion is not:
 - The process is stable.
- It is:
 - The Grid no longer possesses sufficient observational resolution to establish whether the process is stable.
- This distinction should become fundamental to GUDIYA.

Absence of Observed Instability

- The Grid must never mistake blindness for calm.

10. GUDIYA Has Two Responses to Under sampling

- Once the Grid recognizes that evolving cognition is approaching its observational limit, it has two obvious responses.
- It can increase:

$$V_O$$

through greater SPU capacity, faster telemetry, additional observation channels, improved instrumentation or more localized measurement.

- Or it can decrease:

$$V_{RSI}$$

through pacing or throttling.

- Thus:

$$V_O \uparrow \text{ or } V_{RSI} \downarrow$$

until adequate observability is restored.

- This gives ‘Pacing the Frontier’ a much more rigorous meaning.
- Sometimes GUDIYA slows cognitive evolution not because it has detected danger.
- It slows evolution because:
 - The Grid is approaching the limit at which it can reliably tell whether danger exists.

11. Benchmarks Are Sensors

- The same logic applies to evaluation benchmarks.
- A benchmark is not merely a scoreboard.
- From a control perspective, it is a sensor.
- If a benchmark saturates, its ability to discriminate between successive generations declines.
- Thus:

$$\boxed{\text{Benchmark Saturation} \rightarrow \text{Loss of Measurement Information}}$$

- A system moving from:

$$99.90\% \rightarrow 99.95\% \rightarrow 99.97\%$$

may appear to be demonstrating continuing improvement.

- But if the benchmark no longer meaningfully discriminates among capabilities, the measurement channel itself has degraded.
- A saturated benchmark should therefore sometimes trigger:
 - Our sensor may have gone dead.

rather than:
We have solved intelligence.

12. Ground Truth Has Its Own Clock

- The problem becomes even deeper with out-of-distribution behavior.
- Benchmarks can be run quickly.
- Simulation can be accelerated.
- Internal evaluation can operate at machine speed.
- But some consequences of an improved intelligence become observable only through interaction with the external world.
- That world possesses its own timescales.
- Consequently:

Compute Speed $\neq$ Ground-Truth Generation Speed

- An intelligence may be capable of producing another generation in minutes while trustworthy evidence about the consequences of the previous generation requires hours, days or considerably longer.
- No amount of recursive computation automatically compresses every external causal process.
- The Evidence Clock therefore imposes a constraint that additional intelligence alone cannot necessarily eliminate.

13. Passing the Evidence Test Is Still Not Enough

- Now assume that the observation problem has been solved.
- Assume I_{n+1} has been adequately observed.
- And assume reliable evidence establishes:
$$I_{n+1} > I_n$$
- Have we demonstrated that deploying I_{n+1} improves the larger system?
- No.
- This is where GUDIYA introduces another independent constraint.

Actant Improvement $\nRightarrow$ System Stability

- The model can become demonstrably better while the Cognitive Field containing it becomes less stable.

14. The Cognitive Field Changes When the Model Changes

- An improved actant does not exist in isolation.
- It operates inside:

$$\mathcal{F}(t) = \{I_n, A_1, A_2, \dots, A_k, C(t), E(t)\}$$

where $C(t)$ represents changing couplings and $E(t)$ the surrounding environment.

- Replace I_n with I_{n+1} , and the system itself changes:

$$\mathcal{F}_n(t) \rightarrow \mathcal{F}_{n+1}(t)$$

- The new model may reason faster.
- It may initiate more Rendezvous.
- It may coordinate more agents.
- Its persistence may increase.
- Its recursion depth may change.
- Its latency may decrease.
- Its coupling gain may increase.

- Its influence may propagate farther before slower actants respond.
- Every benchmark could improve while the surrounding Cognitive Field loses stability margin.

15. A Better Component Can Produce a Worse System

Engineering provides countless analogues.

- A more powerful engine can overwhelm a drivetrain.
- A faster controller can destabilize a slower plant.
- A more sensitive sensor can amplify noise.
- A faster trading algorithm can alter market feedback.
- A more capable AI can alter the dynamics of an agentic ecosystem.
- Thus the important transition is not merely:

$$I_n \rightarrow I_{n+1}$$

- It is:

$$(\mathcal{F}, I_n) \rightarrow (\mathcal{F}', I_{n+1})$$

- The object requiring stability analysis is therefore the resulting coupled system, not merely the improved model.

16. The Stability Limit

- Let:

$$V_S$$

represent the effective rate at which GUDIYA can detect, characterize and stabilize the systemic consequences of evolving cognition.

- Then another conceptual condition appears:

$$V_{RSI} \leq V_S$$

- If the intelligence changes faster than the Grid can absorb and stabilize the consequences of those changes, successive generations can outrun the stability infrastructure.
- This is the Stability Limit.

17. The Three Clocks

- We can now see three clocks governing recursive self-improvement.

- The Observation Clock

- Can the evolution be resolved?

$$V_{RSI} \leq V_O$$

- The Evidence Clock

- Can trustworthy evidence establish that the modification constitutes improvement?

$$V_{RSI} \leq V_E$$

- The Stability Clock

- Can the resulting systemic dynamics be stabilized?

$$V_{RSI} \leq V_S$$

- Together:

$$V_{RSI} \leq \min(V_O, V_E, V_S)$$

- Again, these are conceptual rates until operational units and measurement procedures are defined.
- But the architectural proposition is powerful:
 - The fastest safe RSI loop is constrained by its slowest trustworthy external feedback loop.

18. The Limits Are Sequentially Dependent

- These three limits are not merely independent boxes.
- There is a logical sequence:

Evolution → Observation → Evidence → Stability Decision → Next Evolution

- Observation precedes trustworthy evidence.
- Evidence informs whether a modification worked.
- Field observation determines what that modification did to the larger system.
- Stabilization determines whether the resulting dynamics remain acceptable.
- Only then should unrestricted progression continue.
- This turns RSI from an uncontrolled recursive loop into a closed-loop evolutionary process.

19. A New Kind of Cognitive Aliasing

The Nyquist connection suggests an additional phenomenon worth naming:

Recursive Cognitive Aliasing

Recursive Cognitive Aliasing occurs when successive meaningful evolutionary states occur faster than the observing system can adequately resolve them, causing the observer to infer a trajectory materially different from the one actually traversed.

- Thus:

$$I_1 \rightarrow I_2 \rightarrow I_3 \rightarrow I_4 \rightarrow I_5$$

may observationally appear as:

$$I_1 \rightarrow I_5$$

- The problem is not merely missing information.
- It is misinterpreting dynamics because of missing temporal information.
- That distinction could become important for future RSI evaluation systems.

20. DSMCA Provides the Stability Instrumentation

- Dynamic Stability Margin Consumption Analysis gives GUDIYA a mechanism for watching the second half of this problem.
- For an actant:

$$M_i(t + \Delta t) = M_i(t) - C_i(t) + I_i(t)$$

where:

- $M_i(t)$ is remaining stability margin;
- $C_i(t)$ is stability-margin consumption;
- $I_i(t)$ is Grid stability infusion.

- The Grid also watches:

$$\dot{M}_i$$

and:

$$\ddot{M}_i$$

- Suppose I_{48} is objectively better than I_{47} .
- Limit One is satisfied.
- But after deployment:

$$\dot{M}_A < 0, \dot{M}_B < 0, \dot{M}_C < 0$$

while coupling density increases.

- The actant may be performing beautifully.
- The system may nevertheless be approaching instability.
- GUDIYA therefore says:

$$\boxed{\text{Improved Intelligence} + \text{Declining System Margin} = \text{Pace the Evolution}}$$

21. The Permissible RSI Velocity Can Be Dynamic

- This does not imply a universal fixed speed limit on intelligence.
- Different systems may possess radically different observation, evaluation and stabilization capabilities.
- Therefore:

$$V_{\text{Permitted RSI}}(t)$$

can itself be dynamic.

- Conceptually:

$$\boxed{V_{\text{Permitted RSI}}(t) = f(V_O, V_E, V_S, M, \dot{M}, \ddot{M}, GR, C(t))}$$

where GR represents Grid responsiveness and C(t) the current coupling condition.

- When margins are healthy and observation capacity is abundant:

$$V_{\text{Permitted RSI}} \uparrow$$

- When observational resolution deteriorates or stability margin is being rapidly consumed:

$$V_{\text{Permitted RSI}} \downarrow$$

- GUDIYA therefore does not need a permanent cognitive speed limit.
- It needs an adaptive stability-limited velocity.

22. This Changes “Pacing the Frontier”

- Pacing the Frontier now acquires a much deeper technical interpretation.
- The Frontier should not advance simply at the rate at which laboratories can create more capable intelligence.
- Nor should it be arbitrarily frozen because future intelligence is frightening.
- Instead:

$$\boxed{\begin{array}{l} \text{Capability Velocity} \leftrightarrow \text{Observation Capacity} \leftrightarrow \\ \text{Evidence Capacity} \leftrightarrow \text{Stabilization Capacity} \end{array}}$$

- The Grid can permit faster evolution as humanity builds better infrastructure for observing, validating and stabilizing it.
- Thus the objective is not:
 - Make intelligence slow.
- It is:
 - Make safe intelligence capable of moving faster.

23. The GUDIYA Supremacy Connection

- This also strengthens the GUDIYA Supremacy Principle.
- Every evolutionary child must remain Grid-responsive:

$$\boxed{GR(I_n) \geq GR_{\min} \forall n}$$

- Why?
 - Because without Grid responsiveness, the Observation Clock itself may fail.
 - The Grid may lose sufficient telemetry.
 - Rendezvous may become opaque.
 - Dynamic Stability Ledgers may become unreliable.

- Couplings may become invisible.
- The Grid may no longer possess the information required to establish either evidence or stability.
- Thus the forbidden transition is not:

$$I_n \rightarrow I_{n+1}$$
- It is:

$$(I_n, GR = 1) \rightarrow (I_{n+1}, GR = 0)$$
- At that point the Grid does not necessarily possess evidence that the actant is dangerous.
- It possesses something more immediately actionable:
 - The Grid can no longer establish that the actant remains safely stabilizable.
 - Under GUDIYA Supremacy, that is sufficient reason to withdraw unrestricted Rendezvous privileges.

24. The External Observer Becomes Essential

- There is a deeper systems-science lesson.
- A self-improving intelligence may become extraordinarily good at understanding itself.
- But:

$$\text{Self-Knowledge} \neq \text{Field Awareness}$$
- An actant exists inside a larger environment whose topology, participants, couplings and states may be changing simultaneously.
- Other actants may themselves be adapting.
- Consequently, the evolving intelligence cannot necessarily serve as the sole observer, modifier and stability authority for the system containing it.
- The Grid provides the external reference.
- This connects directly with GUDIYA's earlier work on the Cognition Coordinate System and relative motion.
- An aircraft may know everything about its engine while still requiring external references to determine its motion relative to the atmosphere, terrain and other aircraft.
- Likewise, an intelligence can understand its internal improvement while remaining unable to independently establish the stability of the entire Cognitive Field surrounding it.

25. Three Different Failure Modes

- The synthesis gives GUDIYA useful diagnostic categories.

• Observation Deficit

$$V_{RSI} > V_0$$

The evolutionary dynamics are no longer adequately resolved.

Potential consequence:

$$\text{Recursive Cognitive Aliasing}$$

• Evidence Deficit

$$V_{RSI} > V_E$$

Evolution occurs faster than trustworthy evidence can establish improvement.

Potential consequence:

$$\text{Modification Mistaken for Improvement}$$

• Stabilization Deficit

$$V_{RSI} > V_S$$

Evolution occurs faster than the Cognitive Field can safely absorb its changing dynamics.

Potential consequence:

Improved Actant, Destabilized System

- These failures are related but not identical.
- They require different remedies.

26. The Recursive Improvement Stability Principle

The synthesis can now be formalized as a Cognitome principle.

Recursive Improvement Stability Principle

A self-improving actant may evolve only at a rate supportable by the external feedback systems required to observe its meaningful state transitions, generate trustworthy evidence that modifications constitute improvement, and dynamically stabilize the consequences of those improvements within the Cognitive Field. Passing any one of these tests does not substitute for the others.

Thus:

$$V_{RSI} \leq \min (V_{\text{Observation}}' V_{\text{Reliable Evidence}}' V_{\text{Grid Stabilization}})$$

with the additional constitutional requirement:

$$GR(I_n) \geq GR_{\min}$$

for every operational generation.

27. The Nyquist Lesson in One Sentence

- The Nyquist insight gives GUDIYA one particularly important warning:
 - The Grid must never confuse an unobserved excursion with an excursion that never occurred.
- That principle reaches far beyond recursive self-improvement.
 - It applies to agent interactions, persistent objectives, Rendezvous, Cognitive Field oscillations and potentially any machine-speed phenomenon being observed by a slower observer.
 - The observer's temporal resolution becomes part of the safety architecture.

28. A Different View of the Intelligence Explosion

- The classic intelligence-explosion picture looks like:

$$I_1 \rightarrow I_2 \rightarrow I_4 \rightarrow I_8 \rightarrow I_{16} \rightarrow \infty$$

- GUDIYA sees something closer to:

$$I_1 \xrightarrow{\text{Observe}} I_2 \xrightarrow{\text{Validate}} I_3 \xrightarrow{\text{Stabilize}} I_4 \rightarrow \dots$$

with continuous Grid observation occurring throughout rather than only at discrete checkpoints.

- The improvement loop is therefore embedded inside a larger control loop:

$$\begin{aligned} &\text{Improve} \rightarrow \text{Observe} \rightarrow \text{Validate} \rightarrow \text{Measure Field Response} \rightarrow \\ &\text{Stabilize} \rightarrow \text{Authorize Continued Evolution} \rightarrow \mathcal{U} \end{aligned}$$

- The intelligence can still advance extremely rapidly.
- But its safe operating velocity depends upon the infrastructure surrounding it.

29. GUDIYA Is Not the Brake — GUDIYA Is the Suspension

- This produces a useful conceptual correction.

- The Grid should not be thought of simply as the brake pedal on recursive intelligence.
- A sophisticated suspension system allows a vehicle to travel faster, not merely slower, because it maintains stability as conditions change.
- Likewise, improved observation, evidence generation and stabilization infrastructure can increase:

$$V_{\text{Permitted RSI}}$$

rather than merely constrain it.

- That is consistent with GUDIYA's broader philosophy:

Stabilize, not suppress

- The objective is not to prevent intelligence from becoming extraordinary.
- The objective is to create an infrastructure capable of safely accompanying extraordinary intelligence.

30. Conclusion — Intelligence Cannot Outrun Its Feedback

- Recursive self-improvement does not occur outside systems science.
- It remains an adaptive process embedded in a larger adaptive environment.
- That process encounters three increasingly fundamental constraints:

Evolution cannot safely outrun Observation

because under sampling can create false reconstruction.

Evolution cannot safely outrun Evidence

because modification cannot be assumed to constitute improvement.

and:

Evolution cannot safely outrun Stabilization

because an improved actant can still destabilize the system containing it.

- Together:

Safe RSI Velocity

- This substantially changes the framing of recursive self-improvement.
- The relevant question is no longer simply:
 - *How quickly can intelligence improve itself?*
- It becomes:
 - *How quickly can an external system adequately observe that evolution, establish that it truly constitutes improvement, and maintain sufficient dynamic stability as the resulting intelligence enters the Cognitive Field?*
- GUDIYA therefore does not stand before self-improving intelligence demanding that it stop.
- It runs alongside it.
- When observation is adequate, evidence is trustworthy, stability margins remain healthy and the actant remains Grid-responsive:

Proceed.

- When observation begins falling behind:

Increase observation bandwidth – or reduce evolutionary velocity

- When evidence falls behind:

Do not confuse modification with improvement

- When stabilization falls behind:
 Pace the Frontier
- And when Grid responsiveness disappears:
 Deny Rendezvous and contain
- The resulting GUDIYA proposition is therefore not anti-intelligence and not anti-self-improvement.
- It is a proposition about the physics of feedback:
- Self-improving intelligence is welcome.
- But evolution cannot outrun observation.
- Observation must be sufficient to produce trustworthy evidence.
- And evolution cannot outrun our capacity to stabilize what that intelligence becomes.

Observe fast enough to know. Know enough to validate. Stabilize fast enough to proceed

