Can Intelligence Grow from Within?

Zonghuan Xu · September 2026

On recursive self-improvement, evolving evaluators, and whether intelligence can find its own direction.

Recently, while reading work on recursive self-improvement (RSI) and self-evolving systems, I have kept noticing a gap between what these ideas lead us to expect and what the evidence actually shows.

“Self-improvement” readily brings to mind a complete cycle: a system identifies its limitations, proposes improvements, tests and incorporates them, then uses its stronger capabilities to carry out the next round of research. What makes this exciting is that the ability to improve is itself part of the loop.

What is demonstrated, however, is often only one part of that cycle. Some systems accumulate experience, update workflows, or learn when to apply particular strategies. These are useful capabilities. My question is what still separates them from the open-ended self-improvement we imagine.

It would be unfair to dismiss all this work as automated checklist writing. The Darwin Gödel Machine, for example, modifies an agent’s own code and selects improvements through coding evaluations. Absolute Zero Reasoner generates and solves its own tasks, using a code executor to provide verification signals. These systems genuinely take over parts of the improvement process that would otherwise be performed by humans. Darwin Gödel Machine; Absolute Zero.

But these results leave a question open: **a system can become better at passing an evaluation without becoming better at recognizing what that evaluation leaves out.**

It is the distance between these two abilities that interests me.

When using AI for research, it is easy to feel that we are almost there. Once a problem has been clearly formulated, many concrete tasks seem possible to delegate: derivations, programming, experimental design, error checking, and organizing results. Humans, meanwhile, still often have to decide what to investigate, which anomalies deserve attention, and when to abandon the original formulation of a problem.

This difficulty is hard to measure in terms of workload. Automating most of the execution does not necessarily mean that most of research autonomy has been achieved. The small amount of work involved in deciding what to do next may determine whether all the remaining work is meaningful.

Of course, choosing a research direction need not depend on some uncomputable intuition. It might be learned, or even explicitly articulated. The difficulty is that its value is usually harder to verify promptly than success or failure on a local task. The value of a new concept may only become apparent after it connects problems that previously seemed unrelated.

This brings the discussion to the evaluator: the mechanism by which a system judges what counts as an improvement.

If the evaluation standard is fixed, a system may keep discovering shortcuts within it. If the system is allowed to change the standard, a different difficulty arises: is it correcting a poor proxy, or redefining success to make success easier?

Suppose a research system notices that its current scoring rule favors statements that are easy to prove but unimportant, and begins rewarding concepts that connect different fields instead. That looks like an improvement in research judgment. But if it starts rewarding grand-sounding statements with little explanatory power, it could just as easily claim to have discovered a more sophisticated standard.

Allowing the evaluator to evolve therefore does not, by itself, solve the problem. We still need an account of why the new way of evaluating deserves greater trust.

This does not mean that we must find an eternal function that assigns a score to every possible activity. It helps to distinguish our underlying concerns, our understanding of those concerns, and the particular indicators used to measure them. Someone can continue to care about health while revising both their understanding of health and the ways they measure it. Relatively stable concerns are compatible with evolving evaluation mechanisms.

A fixed goal is also conceptually compatible with recursive self-improvement. Schmidhuber's Gödel Machine envisages a system that uses formal proofs to decide whether to rewrite itself in pursuit of a predefined utility objective, even rewriting the methods by which it searches for improvements. It shows how justified self-modification can be defined, but does not guarantee that such improvement will continue in practice, nor explain where the original objective should come from. Gödel Machines.

What I really care about, then, is whether a system can recognize when a local standard has failed and find a justified way to revise it.

Mathematics makes this question especially interesting.

Mathematics seems to offer the possibility of progress driven largely by internal activity. Given definitions and rules, a system can discover previously unknown structures through reasoning, constructions, counterexamples, and computational experiments. It may become more capable even without new observations of the physical world.

Here, "no new input" does not mean "nothing new gained." For an agent with limited computational resources, a conclusion being logically contained in its premises is very different from the agent being able to find, understand, and use that conclusion. A good proof or representation can make a previously intractable problem tractable.

This makes me reluctant to accept the absolute claim that intelligence must continually engage with the external world in order to improve. At least for some capabilities, internal reasoning can bring substantive progress.

But mathematics also makes another difficulty particularly clear: proving a statement and judging that it is worth studying are different things.

Within a formal system, we can have clear constraints on correctness while still facing an enormous space of choices. Why study this definition? Why is one generalization more natural than another? Why does one theorem remain an isolated fact while another changes how we organize an entire field?

Must these judgments come from human communities? I am not sure. Perhaps mathematical structure itself can supply some direction: greater unification, wider applicability, lower description costs, and more efficient subsequent reasoning.

Compression and learning progress also have a long research history as sources of intrinsic motivation. Schmidhuber, 2010.

But even if these directions can be formalized, questions remain. Compress what? Improve reasoning efficiency for which problems? Why should these structural gains transfer to broader capabilities? The possibility of intrinsic standards cannot simply be treated as proof that they can sustain unlimited self-improvement.

We can therefore imagine two approaches with different emphases.

One starts from real-world problems and human needs, letting a system improve through ongoing feedback. The other relies more heavily on mathematics, abstraction, self-play, or compression to develop general capabilities first, and then connects those capabilities to the world.

The second approach is attractive. If a system has grasped sufficiently deep general structures, might understanding the human world require only a small amount of additional learning?

I am willing to take this as an important hypothesis, but two steps cannot be skipped. First, does progress in a formal world transfer broadly? Second, how much information about the real world does that transfer require?

No matter how powerful its reasoning, a system cannot, from the same internal state alone, distinguish between two real-world situations that are both consistent with its evidence but require different actions. Better reasoning may help it make more effective use of an observation, but cannot guarantee that the observation becomes unnecessary. Similarly, understanding what a person wants and deciding how to respond to those wishes are different questions.

Simulating people may provide abundant, inexpensive feedback. But the extent to which that feedback represents real human needs still requires calibration against reality. Simulation can amplify existing understanding; it can also amplify existing misunderstandings.

I remain excited about RSI. But beyond asking whether a system can modify itself, I now want to ask:

When it begins to change how it judges progress, how can it keep discovering that it is wrong?