MIT TECHNOLOGY REVIEW KOREA
Interview | English Edition

The Age of Self-Evolving AI

> An era has begun in which AI rewrites its own code, without human direction, to improve its performance. AI builds better AI, and the better AI improves itself in turn. The moment this loop closes, Jeff Clune says, is the singularity: an event horizon past which you cannot see.

By Sooyeon Park, Senior Editor

Jeff Clune | Professor of Computer Science, University of British Columbia · Co-founder, Recursive Superintelligence

One by one, the tasks of building AI are slipping out of human hands. AI is already refining code and finding better ways to train models. Now it has reached the point of improving itself. The Darwin Gödel Machine[1], a system that rewrites its own code, lifted its score on a benchmark of real-world software engineering problems from 20 percent to 50 percent through repeated self-modification alone, with no human direction. Along the way it devised improvements no one had asked for: better code-editing tools, and a scheme for generating multiple candidate answers and letting another AI judge among them.

Industry is moving in the same direction. AI builds a better AI; the better AI then improves itself again. Once that loop closes, AI no longer waits for people to fix it — it carries its own progress forward. The major AI laboratories have thrown themselves into research on closing the loop, and the startups pursuing it are drawing capital and talent alike. Until now, the race has been about pouring resources into ever-smarter models. The next race will be judged by a broader measure: how much discovery and invention those resources can generate.

How far, then, have systems that use AI to build better AI actually progressed? And where does the human role end?

Jeff Clune has worked on AI that builds AI from the beginning. In his 2019 paper on “AI-generating algorithms” (AI-GAs), he laid out a systematic research paradigm in which AI produces better AI, rather than humans hand-designing every component. After stints at OpenAI and Google DeepMind, he led the Darwin Gödel Machine research at the University of British Columbia, and early this year he co-

founded the startup Recursive Superintelligence to scale the vision into working systems. We asked him how far these systems have come, what is driving their progress — and where they must stop.

How Foundation Models Accelerated the Age of AI Building AI

Q. Your 2019 proposal — hand the design of AI over to AI itself — and the LLM boom of the past seven years looked like divergent paths. Yet the Darwin Gödel Machine runs on those very models as its engine. How did the two roads converge?

Foundation models — the broader class of large models that includes LLMs — have become the Great Catalyst of AI-GAs. They have accelerated everything immeasurably, which is exactly in line with the AI-GA philosophy: AI improving AI. What I proposed in 2019 was the idea of AI-generating algorithms, or AI-GAs — instead of humans hand-designing every component of intelligence, we hand the design work over to AI itself.

On reflection, that should not surprise anyone. To borrow from Newton, foundation models allow AI-GAs to stand on the shoulders of giant human datasets. Because they learned to code and think from training on the internet, they can be used to design better AI agents. We showed that, and took the idea further with the Darwin Gödel Machine.

Foundation models also allow AI to select the next interesting task to learn, and to generate an endless supply of training challenges. Genie, a project I worked on at DeepMind, is one example: we trained a world model on internet-scale video so that a neural network can dream up any possible environment — generating an interactive, video-game-like training world from any request. Interestingly, I predicted this would be possible in the AI-GA paper, but thought it was a crazy idea that might take decades. It only took five years to build. And only seven years later, the hottest topic in AI is recursive self-improvement, or RSI — which is another name for AI-generating algorithms.

Q. You have watched AI revise its own code from closer range than almost anyone. What was the moment that most surprised you?

I think the most surprising moment was not a specific event within any one run, but the overall realization that a new type of reinforcement learning is now possible. I have not publicly written these thoughts up yet, but I call it Intelligent Reinforcement Learning.

The standard paradigm has been to rely on reward functions: try many things, and prefer whatever scores well. But strengths and weaknesses are not always reflected in higher or lower scores. Now, like humans, we can ask agents to reason about what is and is not working — even when those strengths and weaknesses are not reflected in the score — and improve themselves through their own intelligent reasoning and judgment. That is an amazing new possibility. It means agents can learn much more sample-efficiently, like humans do: reflecting on what is and is not working and making changes accordingly, without necessarily being guided by a reward function.

Q. You must also have seen where self-improvement stalls. What are the limits today?

Unfortunately, there are still many areas where models do not perform as well as we would like. They are not as creative as humans, nor are they as sample-efficient. They can make mistakes in reasoning. The biggest problem is that they are very susceptible to reward hacking — gaming the score rather than truly achieving the goal. At Recursive, we are working to address all of these limitations, and we are making great progress.

The Failed Versions Behind the Best Result

Q. Your system kept every version of itself, even the ones whose scores got worse — and the best result on record came by way of those degraded versions. Why keep the lower-performing versions?

There is a paradox in life: if you try too hard to solve a very ambitious challenge, you will fail. That is because we do not know how to reward progress toward very ambitious goals. For example, if you went back in time and only funded scientists who improved the ability to cook quickly without smoke, the modern microwave would never have been invented. Its invention required someone working on radar to notice a chocolate bar melting in his pocket.

Similarly, if from the age of the abacus you had only rewarded faster calculation, the computer would never have been invented — it required research into electricity and vacuum tubes, neither of which was developed to improve computation. Instead, we must do what the scientific community does, and what humanity has done since the dawn of time: be curious, explore, and make new inventions and discoveries because they are interesting, until eventually we hold the building blocks that solve the original ambitious problem.

The same is true when searching for ever-better AI. Because we do not know exactly what to reward, we must keep expanding an archive of interesting, diverse stepping stones — even the ones that score poorly for now. In practice, finite compute forces a trade-off between exploration and exploitation, and learning to balance the two takes experience. But for truly hard problems, exploration is by definition critical — and most people's instincts are tilted far too much toward exploitative, greedy search.

Q. The AI Scientist², which automates the entire research pipeline from idea to experiment to finished manuscript, produced a paper that cleared the review bar at a workshop of a leading AI conference. You have read its output yourself. Was there an idea that impressed you as a researcher?

Yes. I read one of its papers and thought the core idea was simple and clever. I thought to myself: if a graduate student came to me and proposed this idea, I would say, "That is a good idea — go try it." For the record, I very rarely say that to students. And when I checked the literature, no one had proposed the idea yet. Then, remarkably, just a few weeks later, a team of human researchers at Oxford published a paper exploring essentially the same idea. Human scientists often invent things independently at around the same time — evolution by Darwin and Wallace, calculus by Newton and Leibniz.

But this was the first time I had seen an AI and humans, standing on the shoulders of the same giants, independently come up with the same non-trivial idea. I think it was a historically significant moment. I should point out that the human team explored the idea far, far better — their paper is vastly superior, and The AI Scientist's paper is not great. But it did come up with the same clever idea, so on the question of whether AI scientists can be truly creative, it showed that they can.

Q. Once AI begins pouring out papers, who does the sorting?

As in many other areas of life, we will need the help of AI judges and evaluators to manage the flood of AI-generated content.

"The Line Where We Must Stop Cannot Yet Be Written Down"

Q. You have moved through academia, OpenAI, and DeepMind. Why carry this research forward in a startup, rather than a university or Big Tech?

We founded Recursive because we believe it is now time to scale these ideas up. We actually think we are at the beginning of a new scaling paradigm, in which we will create algorithms that produce more discovery and innovation as they receive more compute. Scaling these ideas requires tremendous resources, and offers tremendous economic value. A company is therefore the right place to pursue those opportunities.

That said, there is still a lot of important research that can be done in more traditional settings — a university lab, or a small group. I will not yet say which activities will be automated first. But one thing is clear: automation should never eliminate human oversight of, and responsibility for, alignment — keeping AI aligned with human goals and values. We want systems that pursue human goals, including human flourishing, and that keep those who oversee them informed about what they are doing.

Q. You warned about both the promise and the peril of self-improving AI earlier than almost anyone. How long can this research go on?

Let me share my personal views here; they do not represent the position of any organization I am affiliated with. I expect the process to be iterative rather than a single decision. As we proceed, we may encounter worrying signs. If so, we stop and work to eliminate the root cause; once the problem is satisfactorily resolved, we continue. The technology is changing too quickly and too unpredictably to define specific red lines today that would be practically useful.

What matters most is that everyone doing this work uses their best judgment to proceed safely. Complicating matters, no organization is the only actor. Even if some stop at a red line, others will race past it. To me, that is a compelling argument that the best people need to figure out how to proceed safely, rather than standing on the sidelines while less safe or less scrupulous actors race ahead. Ideally, wise regulation and regulators would help coordinate all actors — but sadly, I am not optimistic that such coordination will materialize.

Q. This research requires a staggering amount of compute. A single evaluation run of the Darwin Gödel Machine costs roughly $22,000. Is there room in this race for countries outside the frontier labs, such as Korea?

I am not an expert on Korea's infrastructure or on the best strategy for Korea going forward, so I cannot comment on those specifics. More broadly, there are tremendous opportunities across AI. There is much research to be done, some of it possible on modest budgets. There are many opportunities to apply AI to build useful products. One can also build on the many very impressive open-source models. I think humanity has just discovered a vast orchard of low-hanging fruit, and opportunities abound.

Q. We have entered an era in which even research is done by AI. Where should researchers just entering the field place their focus?

Until the singularity — when true recursive self-improvement takes off — I think there will be an endless supply of important research questions to pursue. One easy way to find them: look at human and animal brains, identify the dimensions along which they outperform current AI systems, and try to close the gap. Natural brains are far better than today's AI at learning continually, generalizing out of distribution — coping with situations they have never seen — learning from few examples, and being energy efficient. There will soon be major advances in all of these areas, and many of them will be made by students entering the field this year.

Critically, we also need as much AI safety and alignment research as possible, to raise the probability that this technology develops in ways that benefit humanity. And AI can be applied to countless domains, from cancer research to cleaner energy. Researchers should not be discouraged. They should be

energized by the fact that they can wield this new technology to make rapid progress. After the singularity, it is hard to predict what will happen — that is why it is called a singularity: an event horizon past which you cannot see. All aspects of society may be reimagined, but that applies equally to all areas of the economy, not just research.

---

[1] Darwin Gödel Machine: A self-improving AI unveiled in 2025 by Professor Clune's research group together with Japan's Sakana AI. Rather than discarding versions of itself whose scores decline during self-modification, it archives them and branches off in different directions from the versions it has kept.

[2] The AI Scientist: An AI system that automates the entire research process, from generating ideas and running experiments to writing the manuscript and reviewing its own work. One paper it produced received review scores above the acceptance threshold at a workshop of a leading AI conference.

Jeff Clune was a professor at the University of Wyoming and a senior research scientist at Uber AI Labs before leading a research team at OpenAI and serving as a senior research advisor at Google DeepMind. He is a professor of computer science at the University of British Columbia and a Canada CIFAR AI Chair at the Vector Institute, and in early 2026 he co-founded Recursive Superintelligence.