SDS PODCAST

EPISODE 1004: Recursive Self-Improvement

Jon Krohn: 00:00 This is episode number 1004 on recursive self-improvement. Welcome back to the SuperDataScience Podcast. I'm your host, Jon Krohn. Today's topic is Recursive Self-Improvement or RSI for short. This is the idea that an AI system could get good enough at AI research to build its own more capable successor, which then builds an even better one and so on and so on and so on in a loop that compounds with every turn. The phrase RSI exploded across the AI world this month after Anthropic published a report on its own internal engineering. The headline number is striking. As of May 2026, more than 80% of the code merged into Anthropic's production code-based was written by Claude, the company's AI model. They also sponsor this podcast and this episode, but they don't have any editorial input. I decided to do this topic myself and all the research was done independently.

01:03 So yeah, if you hear a Claude ad, it has nothing to do with the substance in this episode. Anyway, before Claude's coding agent launched in early 2025, instead of the kind of 80% figure that Anthropic is reporting in May of this year, the figure last year was in the low single digits. The typical anthropic engineer is now shipping roughly eight times as much code per day as they were two years ago, not by working harder, but by directing an AI that does the typing. In one case this past April, Anthropic said Claude ship more than 800 fixes that cut a whole class of API errors a thousand fold. Work the engineer overseeing it estimated would've taken a human four years. So is this RSI? Is this recurse of self-improvement? Not quite. No. And the distinction matters. What's happening today is AI-assisted coding, where human engineers still set the goals, organize the work, and judge the results.

01:59 True RSI would be when the AI manages the whole endeavor, designing the experiments, writing the code, evaluating what worked and deciding what to try next

with little or no human in loop. The term itself isn't new. The British mathematician IJ Good described an intelligence explosion back in 1965, reasoning that a sufficiently capable machine could redesign itself and rapidly leave human intelligence behind. The concern was later formalized by AI safety researchers and the rise of large language models has dragged RSI out of science fiction and into quarterly engineering reports. The evidence that we're inching toward RSI is accumulating. Researchers at the think tank meter, whom I talk about all the time on this show, have been measuring how long a task an AI agent can complete on its own defined by how long the same task takes a human professional. They found that this time horizon had been doubling roughly every seven months for the past six years, but over the past year it appears to have sped up to something closer to every four months that is staggering.

03:04 In practice, frontier models went from reliably handling software tasks that take humans a few minutes to tackling work that would take a skilled engineer most of a working day. Anthropic's own breakdown tells a similar story. On the hardest, most open-ended problems, where even the definition of success is fuzzy, its model's success rate jumped from under 20% in late 2025 to 76% by May of this year. That's only a few months.

03:32 Two concrete examples bring this wild acceleration to life. In 2025, one of Google DeepMind systems Alpha Evolve started designing algorithms on its own. It found a better way to schedule workloads across Google's data centers that recovered nearly three quarters of a percent of the company's worldwide computing power. And it discovered a faster method for matrix multiplication that sped up the training of Google's flagship model by about 1%. Those are small percentages, a quarter of 1% and 1%, but those are very small percentages on enormous numbers. So the absolute savings are huge and crucially, the AI was improving the very machinery used to build AI. The

second example I have for you is even more pointed. Andrej Karpathy, a founding member of OpenAI's research team and the former head of AI at Tesla released a small tool that lets an AI agent autonomously tune a model training script overnight.

04:22 He pointed it at code he had already carefully optimized himself. Over about two days, the agent ran roughly 700 experiments, kept around 20 real improvements and cut the time to train a GPT quality model from 2.02 hours down to 1.80 hours. That's an 11% speed up on already excellent code and the agent code efficiencies that Karpathy himself had missed. As Karpathy put it, none of the individual tricks were especially novel, but they stacked up and he didn't have to touch a thing. Back to the anthropic report that I was talking about at the outset of this episode. In that report, they laid out three scenarios for where this goes from here. In the first scenario, AI stays below the level of the best human engineers and humans remain firmly in charge. In the second, which Anthropic and I would consider most likely in that second scenario, AI assisted engineering keeps accelerating, but humans still steer model research and development.

05:18 In the third scenario, AI becomes capable of improving itself. Anthropic co-founder Jack Clark has put a 60% probability on an AI system being able to fit into that final self-improving bucket, being able to create its own successor with no human involvement at all by the end of 2028, so a couple of years from now. Now, plenty of smart people think that timeline is too aggressive and their reasons are worth taking seriously. The first bottleneck is compute. Even with efficiency gains, each new generation of models needs more computing power to train. So progress is chained to the pace of data center construction and every chip serving a paying customer is a chip not available for open-ended research. The second bottleneck is data. AI has improved fastest where success

is cheap to verify automatically. Code either runs or it doesn't. A math proof is either valid or it isn't and that lets models safely learn from data they generate themselves.

06:12 It's far murkier to check whether a model has gotten better at creative writing or legal judgment and there's a real risk of what researchers call recursive drift where small errors in a model's own output compound as it trains on itself. There's also a sharper critique, which is that some of the framing is marketing. A market leader calling for the world to have the option to slow down frontier AI development is conveniently also a market leader asking its competitors to ease huff the gas. Several prominent researchers have pointed out that the gap between today's agentic coding and actual RSI is wider than the excitement marketing suggests. But Anthropic's leadership appears sincere in its concern and they're not alone. The physicist Max Tegmark likens racing towards self-improving AI without adequate safeguards to flooring the accelerator on a highway with your eyes closed, fine for a while right up until it very much isn't.

07:07 The worry isn't a Hollywood robot uprising so much as a quieter loss of control. A world where models are trained by models to pursue goals set by models with safety verified only by other models and humans gradually edged out of the decisions that matter. So where does that leave us? I'd land as regular listeners will have come to expect somewhere in the optimistic middle. The productivity gains here are real and already in your hands. The same coding agents, accelerating anthropics engineers are available to you and me right now and they make building useful things dramatically cheaper and faster than they were even a year ago. At the same time, the closer we get to systems that improve themselves, the more it pays to keep our eyes open, to build in monitoring, human checkpoints and oversight while these tools are still firmly under our direction. RSI is

increasingly in the vernacular of our industry and it's certainly a threshold to keep our eyes on.

08:00 If you're concerned about runaway AI systems and would like to do something about it, I highly recommend checking out episode number 1007 of this podcast when it comes out in a couple of weeks. It'll feature Ben Todd explaining the greatest AI threats facing society and how you can get yourself into a career combating these risks. If you can't wait until July 7th when that episode comes out, you can check out Ben's previous appearance on this show from five years ago. That's episode number 497. All right, that's the end of today's episode. If you enjoyed it or know someone who might consider sharing this episode with them, leave a review of the show on your favorite podcasting platform or on YouTube, tag me in a LinkedIn post with your thoughts. And if you aren't already, be sure to subscribe to the show. Most importantly, however, we hope you'll just keep on listening.

08:48 Until next time, keep on rocking it out there and I'm looking forward to enjoying another round of the SuperDataScience Podcast with you very soon.