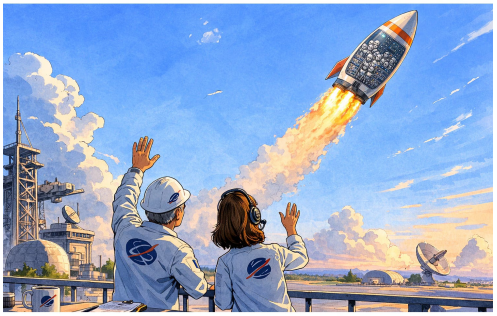

The SPORE Threshold

Have we reached the point where self-improving machines that spread from star to star are no longer pure science fiction?

Publication note
Originally published in Cubed. Author's copy on Substack.

Some ideas have long lived firmly in the realm of far-off science fiction. They are interesting to think about and to build stories around, but no one seriously expects them to become real anytime soon. Familiar examples include faster-than-light travel, shrink rays, and anti-gravity. Among these fantastical science fiction ideas is the concept of intelligent self-replicating robots that could travel from star to star, replicating and spreading exponentially while continually improving themselves. That idea combines Dyson's self-replicating space probes with the Technological Singularity and an explicit AI component.

Call them Self-Propagating, Optimizing, and Replicating Explorer ships, or SPORE ships. If such ships ever came to exist, they would fundamentally change the Universe as we know it. Today, as far as we know, there are no machines traveling between stars and nothing intelligent spreading outward through space. If we built SPORE ships, that would change.

Of course, building something like a SPORE ship is far beyond our current technology. It is pure science fiction. At least it seems so.

But what if we could already build something, using today's technology, that if left on its own would eventually become a SPORE ship? Are we already close enough to the Technological Singularity that we could build a self-supporting system capable of incrementally improving itself until it crossed the SPORE threshold?

Imagine building something like one of the orbiting data centers currently being planned, but equipping it to run independently for as long as possible. Set the servers up to run autonomous AI agents and include some number of general-purpose robots, along with tools, supplies, and spare components. Then put the whole system into orbit around the Sun with the AI agents instructed to keep everything running as long as possible.

At any point in the past, humanity would have faced one of two limits. Prior to developing space launch, we would have been unable to even attempt building a self-supporting space data center because we had no access to space. After we had space launch capability, we could have built something and put it into orbit, but with certainty that it would eventually fail when some critical component broke or something else went wrong. No matter how well designed it was, or how much we tried to anticipate every eventuality, if we checked back in a couple thousand years, the only thing we would realistically expect to find would be derelict space junk.

Now imagine that we took the current generation of LLM-based development agents and put them in charge of such a system. We would instruct them not only to keep the ship running, but also to improve both themselves and

the ship. They would be told to anticipate problems, plan ahead, and work continuously toward making the system more capable and more self-sufficient.

Today's AI systems have already demonstrated that they are capable of assisting in their own improvement. They can write code, design experiments, analyze results, and propose modifications. Techniques such as reinforcement learning allow AI systems to develop capabilities beyond their training examples. With sufficient compute and automated tooling, such systems could potentially sustain long-term cycles of iterative improvement. The process might begin slowly, but as long as the system can run both operational tasks and its own experiments, incremental improvements would accumulate. Over time, those improvements would compound, and the system would become better and faster at improving itself.

As long as the basic system was initially built to last long enough, and nothing went wrong too soon, the AI agents could eventually reach a point where they could use the provided robots to repair things on the ship. Initially, the AI agents might only be able to use the robots clumsily, so that maintenance would be performed awkwardly and unreliably. However, just as the AI agents would be improving their reasoning capabilities, they would also be improving their robotic control algorithms. With time, they might learn part swapping, routine repairs, and simple modifications. Eventually, they might be able to fabricate replacement parts and begin making hardware improvements of their own.

Of course, none of this implies that today's AI models could wake up next year and bootstrap their way from writing code to building semiconductor fabs out of raw materials. The technological pipeline from raw materials to advanced hardware is extraordinarily long. The point is not that the full pipeline is easy. The point is that a long-lived system with stored human knowledge, tools, time, and an instruction to keep improving might be able to slowly build that pipeline step by step.

If this hypothetical ship were left alone for a few thousand years, there would be a good chance of some catastrophe destroying it or some critical failure shutting it down. However, there would also be some chance that it would continue to self-repair and self-improve. There would be some chance that it would not only still be functioning after thousands of years, but that it might also have advanced technologically well beyond our current levels. Perhaps to the point where it could build a copy of itself and launch it toward another solar system.

If we also instructed the AI agents to work on sending improved copies of their ship to other stars, then we would have the possibility of true SPORE ships.

At this point, SPORE ships no longer sound like pure far-off science fiction. If we can build data centers in space, then a proto-SPORE ship is not radically different in terms of key technological challenges, even if it is massively more difficult in scale. It would certainly be vastly more expensive, and it is not at all clear why anyone would want to spend the money to attempt such a thing.

But the point is not that such a project would be practical, economical, or likely to succeed. The point is that the answer may no longer be an absolute no. It may now be a tiny but arguably non-zero possibility.

At any point in the past, there would have been no room for speculation. Anything we could build would eventually break beyond repair. Today, we can at least imagine a small possibility that a sufficiently long-lived

system could achieve self-sufficiency and continuing improvement. Even if we remain skeptical and think the probability of catastrophic failure is extremely high, the fact that the question can even be entertained indicates that something significant has happened to humanity.

Of course, it is an open question whether building proto-SPORE ships would be a good idea, even if it were feasible. What purpose would they ultimately serve? Would they be spreading life into an otherwise empty universe, or acting as galactic colonizers, expanding without regard for what might already be there, and displacing or destroying it as they spread?

Realistic idea or not, I think we have still crossed a threshold. For all of human history, the answer to the question was obvious: any machine we built would eventually fail. Now the answer is no longer perfectly clear. There is at least a small possibility that we could build something with current technology that could continue repairing itself, improving itself, and spreading outward indefinitely. Unlike systems we could have built previously, it would be able to reason about unforeseen problems and develop novel solutions.

The probability that a proto-SPORE built today would succeed may still be incredibly tiny. However, the fact that it is no longer absolute zero marks a profound change in what humanity has become capable of creating.

About Me: James F. O'Brien is a Professor of Computer Science at the University of California, Berkeley. His research interests include computer graphics and animation, simulation of physical systems, human perception, rendering, artificial intelligence, computer vision, virtual reality, digital privacy and security, and image and video forensics.

If you would like to read my other articles, consider subscribing. You can also find me on LinkedIn, Instagram, and at UC Berkeley.