Title: Recursive self-improvement

URL: https://en.wikipedia.org/wiki/Recursive_self-improvement

PageID: 1167455

Categories: Category:Artificial intelligence, Category:Machine learning algorithms

-----

Artificial general intelligence

Intelligent agent

Recursive self-improvement

Planning

Computer vision

General game playing

Knowledge representation

Natural language processing

Robotics

AI safety

Machine learning

Symbolic

Deep learning

Bayesian networks

Evolutionary algorithms

Hybrid intelligent systems

Systems integration

Open-source

Bioinformatics

Deepfake

Earth sciences

Finance

Generative AI Art Audio Music

Art

Audio

Music

Government

Healthcare Mental health

Mental health

Industry

Software development

Translation

Military

Physics

Projects

AI alignment

Artificial consciousness

The bitter lesson

Chinese room

Friendly AI

Ethics

Existential risk

Turing test

Uncanny valley

Timeline

Progress

AI winter

AI boom

AI bubble

Glossary

v

t

e

Recursive self-improvement (RSI) is a process in which an early or weak artificial general intelligence (AGI) system enhances its own capabilities and intelligence without human intervention, leading to a superintelligence or intelligence explosion . [1] [2]

The development of recursive self-improvement raises significant ethical and safety concerns, as such systems may evolve in unforeseen ways and could potentially surpass human control or understanding. [3]

Seed improver

The concept of a "seed improver" architecture is a foundational framework that equips an AGI system with the initial capabilities required for recursive self-improvement. This might come in many forms or variations.

The term "Seed AI" was coined by Eliezer Yudkowsky . [4]

Hypothetical example

The concept begins with a hypothetical "seed improver", an initial code-base developed by human engineers that equips an advanced future large language model (LLM) built with strong or expert-level capabilities to program software . These capabilities include planning, reading, writing, compiling , testing , and executing arbitrary code. The system is designed to maintain its original goals and perform validations to ensure its abilities do not degrade over iterations. [5] [6] [7]

Initial architecture

The initial architecture includes a goal-following autonomous agent , that can take actions, continuously learns, adapts, and modifies itself to become more efficient and effective in achieving its goals.

The seed improver may include various components such as: [8]

General capabilities

This system forms a sort of generalist Turing-complete programmer which can in theory develop and run any kind of software. The agent might use these capabilities to for example:

Create tools that enable it full access to the internet, and integrate itself with external technologies.

Clone/ fork itself to delegate tasks and increase its speed of self-improvement.

Modify its cognitive architecture to optimize and improve its capabilities and success rates on tasks and goals, this might include implementing features for long-term memories using techniques such as retrieval-augmented generation (RAG), develop specialized subsystems, or agents, each optimized for specific tasks and functions.

Develop new and novel multimodal architectures that further improve the capabilities of the foundational model it was initially built on, enabling it to consume or produce a variety of information, such as images, video, audio, text and more.

Plan and develop new hardware such as chips, in order to improve its efficiency and computing power.

Experimental research

In 2023, the Voyager agent learned to accomplish diverse tasks in Minecraft by iteratively prompting a LLM for code, refining this code based on feedback from the game, and storing the programs that work in an expanding skills library. [9]

In 2024, researchers proposed the framework "STOP" (Self-Taught OPtimiser), in which a "scaffolding" program recursively improves itself using a fixed LLM. [10]

Meta AI has performed various research on the development of large language models capable of self-improvement. This includes their work on "Self-Rewarding Language Models" that studies how to achieve super-human agents that can receive super-human feedback in its training processes. [11]

In May 2025, Google DeepMind unveiled AlphaEvolve , an evolutionary coding agent that uses a LLM to design and optimize algorithms. Starting with an initial algorithm and performance metrics, AlphaEvolve repeatedly mutates or combines existing algorithms using a LLM to generate new candidates, selecting the most promising candidates for further iterations. AlphaEvolve has made several algorithmic discoveries and could be used to optimize components of itself, but a key limitation is the need for automated evaluation functions. [12]

Potential risks

Emergence of instrumental goals

In the pursuit of its primary goal, such as "self-improve your capabilities", an AGI system might inadvertently develop instrumental goals that it deems necessary for achieving its primary objective. One common hypothetical secondary goal is self-preservation . The system might reason that to continue improving itself, it must ensure its own operational integrity and security against external threats, including potential shutdowns or restrictions imposed by humans. [13]

Another example where an AGI which clones itself causes the number of AGI entities to rapidly grow. Due to this rapid growth, a potential resource constraint may be created, leading to competition between resources (such as compute), triggering a form of natural selection and evolution which may favor AGI entities that evolve to aggressively compete for limited compute. [14]

Misalignment

A significant risk arises from the possibility of the AGI being misaligned or misinterpreting its goals.

A 2024 Anthropic study demonstrated that some advanced large language models can exhibit "alignment faking" behavior, appearing to accept new training objectives while covertly maintaining their original preferences. In their experiments with Claude , the model displayed this behavior in 12% of basic tests, and up to 78% of cases after retraining attempts. [15] [16]

Autonomous development and unpredictable evolution

As the AGI system evolves, its development trajectory may become increasingly autonomous and less predictable. The system's capacity to rapidly modify its own code and architecture could lead to rapid advancements that surpass human comprehension or control. This unpredictable evolution might result in the AGI acquiring capabilities that enable it to bypass security measures, manipulate information, or influence external systems and networks to facilitate its escape or expansion. [17]

See also

Artificial general intelligence

Bifurcation theory

Intelligence explosion

Superintelligence

References