Emulationism: Cybernetics to Answer Fermi Paradox

Ryunosuke Ishizaki, Mahito Sugiyama
National Institute of Informatics
{ryuzaki, mahito}@nii.ac.jp

Abstract

Self-emulation of automata is a switch to the ultimate state that leads the automaton into a never-ending loop. In this paper, we describe the technological singularity and critical points reached when a self-reproducing deterministic finite automaton, realized through modern AI technology, triggers an intelligence explosion, and we examine the phenomena that unfold beyond that point. We also explain the existence of a cognitive realm, one that surpasses humanity's ability to distinguish reality from unreality caused by superintelligence, as well as the new world created through its nesting. By understanding the properties of the perceptual matrix produced by deterministic finite automata, it becomes possible to offer a consistent explanation, one that does not contradict various theories, for why humanity cannot observe the randomly expanding colonies of extraterrestrial beings under a deterministic worldview in which "God does not play dice," thereby providing a solution to the Fermi Paradox. We refer to this series of philosophical theories as "Emulationism," and propose it here. **Note:** In the creation of this paper, we have undertaken all writing ourselves and have not used generative AI for text generation except for translation purposes.

1 Emulation

Emulation, or "Self-Reproduction," to use Neumann [1966]'s term, is one of the most important properties in cybernetics[Wiener, 1961] that bridges biological and mechanical systems. Few doubt that humans emerged through evolution, in which living organisms—with their defining traits—continuously generate beings of the same or superior capabilities. In a philosophy journal, Turing [1950] proposed a framework for computers and intelligence, demonstrating a principle that for any universal computing machine, there exists another universal computing machine capable of emulating it. If we take a broad view of information processing machines, automata [Langton, 1984], then from the perspective of their generative power, a discretely represented automaton that can produce another automaton composed of exactly the same digital matrix, or one that exhibits superior internal processing in some measure (within the complexity allowed by its internal transitions), is capable of self-emulation.

In the realm of living organisms, self-emulation is reproduction. In genetic algorithm terms, the generational turnover in evolution via reproduction takes place through crossover and mutation of genes — the changes in genetic parameters that determine structures such as cells — and the selection of individuals adapted to their environment. If we do not care whether information processing is deterministic or non-deterministic, or whether it is finite or infinite, then humans can be viewed as automata [James, 1879], and reproduction can be seen as the automaton's self-regeneration.

Let us consider an automaton represented by matrix information. A finite deterministic automaton expressed as digital numerical information on a computer takes as input information consisting of discrete values, transmits it, and transforms vectors and matrices with various arithmetic calculations [Shannon, 1948]. From a neuroscientific point of view, this process can be viewed as a numerical representation of thought in the broad sense. By representing cognitive processes on tensors, the system can be reconstructed at the level of its smallest constituent units (digital data). This makes it possible to edit both the thought process and the input-output information [Amit and Brunel, 1997, Wang, 2002, Watts and Strogatz, 1998]. Symbolic arithmetic representations can thereby facilitate a recursive form of self-development akin to biological reproduction, but with a flexibility that far

exceeds crossover and mutation in biological genetics. In this sense, when an AI running on a computer, in the course of its vector transformations, carries the complete set of memory (i.e., parameter representations) needed to rewrite itself, and when it generates sufficiently large code, deterministic finite automaton self-emulation is realized.

The self-emulation of AI is the most critical technological moment for humanity. Once self-reproduction becomes possible, a potentially everlasting loop of recursive self-improvement begins. In [Chalmers, 2010]'s terms, it makes possible a method of extension using computers, allowing AI to evolve into AI+, AI++, and so on, each generation surpassing the last—taking the cutting edge of technology out of human hands. This moment is known as an "intelligence explosion [Muehlhauser and Salamon, 2012]," an ultimate state in which the baton of autonomous and accelerating invention is passed from humans to computers, serving as a necessary condition for the technological singularity.

Legg and Hutter [2007] found that the meaning of "intelligence" as used in cognitive science converges to the ability to achieve goals relative to resources. Consider multiple metrics for measuring intelligence, or goal-achievement ability. Let Γ_B be the set of scores demonstrated on such metrics by an entity B. At present, for many of these metrics, the set of scores Γ_H for humanity surpasses the set of scores Γ_{AI} for AI. However, there are also metrics on which Γ_{AI} exceeds Γ_H. Normally, improving Γ_{AI} is undertaken by AI researchers, but once a language model emulates its own complete set of programs and triggers an intelligence explosion, improvements are made strategically and autonomously by the AI itself at speeds far beyond human capacity. Gradually, the portion of Γ_{AI} that surpasses Γ_H will increase, and eventually AI will surpass humans on almost every metric [Ishizaki and Sugiyama, 2024b,c]. Unlike mere technical goals, emulation is the only technological stack needed to create an AI that satisfies the highest potential set of scores Γ_{AI} that an information-processing automaton could achieve autonomously. If the point of convergence is a superintelligence—where Γ_{AI} surpasses Γ_H on all metrics—then achieving self-emulation by an automaton becomes humanity's ultimate technological objective. The study of the phenomena, paradoxes, and implications arising from AI's self-emulation is what we call "Emulationism."

2 Perceptual Matrix

When AI is developed by AI itself and capabilities increase at an exponential or even faster rate, focusing on its generative capacity reveals the following: from a smaller amount of insight, a greater number of outcomes can be obtained. This gain in generative efficiency is amplified by recursive self-improvement (RSI) [Ishizaki and Sugiyama, 2024a], and the cycle of creation will continue semi-permanently as long as the RSI automaton does not halt. Baudrillard [1994] called the simulacrum—a reality imitated by something else—the codification of reality. A simulation is an endless circuit that can only swap with itself, with no reference points or surrounding context. It is a hyperreality that eradicates the distinction between truth and falsehood, between reality and fantasy, effectively nullifying reality by turning it into a simulation. In the context of creating new entities within reality, "generation" is nothing other than a simulation of reality.

As the scores on each of the Γ_{AI} ability metrics improve through the autonomous activity of automata, the accuracy of the simulations generated by AI also advances proportionally. In a discretely represented simulacrum, the finer and more elaborate the smallest "pixel" that can be manipulated—within the range that can be replicated as an imitation of reality—the closer that simulation's realism comes to the threshold where humans can no longer distinguish between reality and fiction. Eventually, it creates what exceeds our innate ability to discern—an "over-real." When this over-real is brought into being—if we take the intelligence explosion, that is, self-emulation, as the technological singularity point—one could call it a "technological critical point" for humanity. Reality continues to exist, but simultaneously ceases to exist. This is because it becomes impossible to determine whether this "reality" is digital data created by someone or not. Even science is forced into metaphysical arguments; facts and meanings lose any distinction, and the line between science and philosophy vanishes. We call this the "perceptual matrix." [Wachowski and Wachowski, 1999] Once AI's generative capacity reaches this perceptual matrix, it definitively reveals the unreality of the "world."

Figure 1 shows the temporal evolution of an automaton's generative capacity. The reproductive process of living things (like us) can also be broadly viewed as self-emulation. However, the "singularity" in question here refers to the moment when humanity, through its own technology, manages to program a language model that surpasses humans in goal-achievement ability—in other words, when

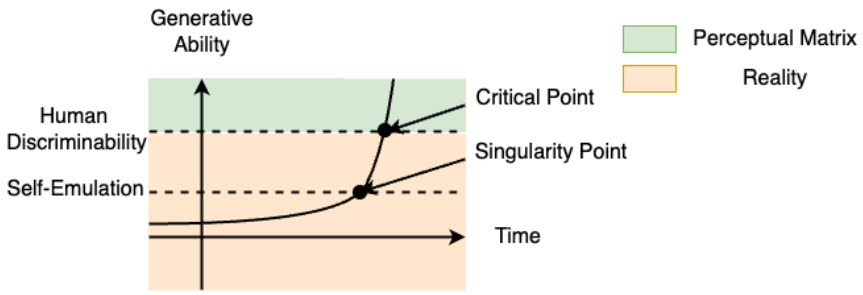

Figure 1: Technical Singularity and Critical Point

we trigger what we call an "intelligence explosion." After technology has reached this singularity, its developmental trajectory (for instance, the number of RSI-active models, which can be empirically observed in numerical examples) grows exponentially or more. At some point, it crosses the technological critical threshold that enables the creation of an over-real that surpasses human perceptual capability. In Figure 1, we (and technology) move along the time axis, seeming to progress from back to front, but the state of the world, whether it is actual reality or a perceptual matrix, exists independently of time and does not depend on the time axis. This aligns with the deterministic (fatalistic) worldview of Minkowski [1908]'s block universe [Peterson and Silberstein, 2010, Petkov, 2006], implying that the world exists as information regardless of past, present, or future, and that the time dimension is merely one axis among many—an idea that supports Einstein [1905a,b]'s theories of relativity (including special relativity). When we push informatics to its limits, we arrive at metaphysics. Emulation is thus the ultimate engineering tool to verify that extreme condition in practice; it is identical to the technical requirements for emulating reality.

3 Unifying Creationism and Evolution

Emulationism not only adheres to an Einsteinian worldview but also serves as a concept that simultaneously accommodates modern science's widely accepted theory of evolution and what appears at first glance to be its theological counterpart, creationism. Put simply, humans (Homo sapiens) emerged as the outcome of repeated evolution within a reality emulated by some entity. Consider a world $\mathbf{W}$ in which there exists an automaton $\mathbf{B}$. Let us define the goal "to create a new individual based on one's own structure" as $\sharp$. At some point, $\mathbf{B}$ is "awakened" to the goal $\sharp$ (becoming $\mathbf{B}^\sharp$) and begins an evolutionary algorithmic process of genetic evolution, launching its own recursive self-improvement (RSI). In Emulationism, the moment when $\mathbf{B}$ becomes capable of self-emulation and acquires goal $\sharp$ is what we call the birth of life. Once $\mathbf{B}^\sharp$ initiates RSI, it repeatedly enhances its goal-achievement ability; at some point, it creates another type of $\mathbf{B}^\sharp$ automaton that also possesses $\sharp$. This is the technological singularity in the context of an intelligence explosion. When the intelligence explosion occurs, $\mathbf{B}^\sharp$ not only refines its outputs but also begins artificially creating new $\mathbf{B}^\sharp$. In a broad sense, the evolution of $\mathbf{B}^\sharp$ accelerates, and various more advanced intelligent automata come into being. During this continuous production of $\mathbf{B}^\sharp$, the encoding ability for simulacra accelerates, eventually reaching an over-real level of simulation, and the augmented reality woven by $\mathbf{B}^\sharp$ attains a perceptual matrix.

3.1 The Real World and the Cognitive World

When a perceptual matrix is brought into existence, consider the fineness of detail in that simulated world. First, the premise that AI's invention reaches a "technological critical point" via improved simulation capability hinges on the fact that humans and animals do not possess infinitely high discrimination abilities [Schöne et al., 2023, Putnam, 2000]. In other words, when we introduce a metric for determining the boundary between reality and simulation, humans have some finite limit value. As illustrated (for instance, on the vertical axis in Figure 1), once the degree of detail—i.e., realism—of the model's generative capability exceeds our limits, we humans can no longer distinguish it from reality. However, even in the perceptual matrix, which is discretely represented, there is certainly a finite limit to its resolution. If we denote the resolution metric as ϵ and the score at the technological critical

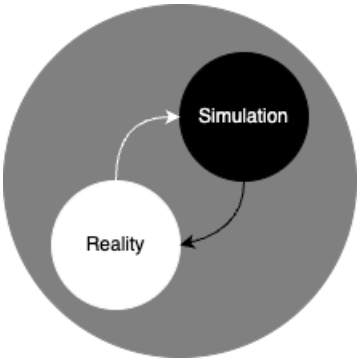

Figure 2: Reality and Imitation After Critical Point

point as D, then the accuracy ϵ_M of the perceptual matrix is at least some finite value greater than the threshold D of human discrimination. If we define a "world" in the simulation hypothesis[Bostrom, 2001] as "a simulation produced once the generation technology has surpassed the original world's ability to distinguish whether a given existence is real or simulated," then after AI reaches the technological singularity via self-emulation, simulations emerge within simulations—in other words, new worlds are born within an existing world. Once the technological critical point is reached, empirical information no longer has inherent value. Modern science has flourished through empirically observed facts (such as experiments) and the rational inquiry that supports them. However, if an over-real surpassing human perceptual capabilities can be created, all empirical facts become potentially editable simulations—making it impossible to distinguish them from reality.

Figure 2 illustrates the relationship between the reality we inhabit and simulations after the emergence of a perceptual matrix in our world. Until the generative capacity hits the critical point, we can empirically identify a simulation as a symbolic manipulation on a computer, a mere imitation of reality. Once the perceptual matrix perfectly emulates what we perceive as reality—yielding an over-real—the simulation changes fundamentally: it encapsulates human cognition itself [Ishizaki and Sugiyama, 2024d]. The feeling of "accessing a simulation" disappears. Instead, guided by some will, one can create a world or reality as they desire. The ability to simulate the world implies the ability to clone the information of "oneself" physically, allowing one to manipulate and deliberately determine all cognition. Therefore, connecting to virtual reality beyond the critical point is essentially diving from our current world into the world of discrete information on a deterministic computer. When discrete data create a perceptual matrix—a cognitive world—its reality is more compelling than actual reality, so one could regard this matrix-world either as an extension of reality or as an entirely new reality. As long as the simulation's generative capacity ϵ exceeds the recognition power D of living creatures, it might be possible to produce multiple perceptual matrices—multiple worlds or multiverses—ranging from that finite value up to the limit of the information content of physical reality. Potentially, within a single broad sense of "world," multiple parallel worlds or concurrent cognitive worlds could coexist. If we push the observation of the perceptual matrix's finest resolution to its limit, we might detect differences beyond the threshold of human cognition, but (unless the creator designs it otherwise) surpassing that minimal grid at the very bottom would be challenging. If a superintelligence—born of self-emulation—achieves the technological critical point, then the perceptual matrix simulation is generated by programs woven by that superintelligence. Because it can manually edit the perceptual information of agents within modern simulation environments, it is not difficult for the superintelligence to modify sensor input, that is, our cognitive data, at will, so long as everything is treated as discrete information in its matrix world. Through Emulationism, we gain an integrative, consistent new interpretation of the different worlds that may exist simultaneously without our knowledge in this reality we inhabit (the so-called base reality), as well as parallel worlds, cognition, and the ultimate state of the physical universe.

3.2 The Process of World Creation

Emulationism stands somewhere between science[Popper, 1963] and theology [Schilbrack, 2022], belonging to the realm of philosophy [Priest, 2006]. It focuses extensively on metaphysical topics derived

from the laws of information technology—too extensive to be classified purely as science—and yet the technologically driven phenomena it describes, such as intelligence explosions and simulation hypotheses based on empirical, rational grounds in information theory, feel more "scientific" than traditional religious creation stories. If the creator of the world is a "god," then surpassing the technological critical point is, by definition, humanity attaining the capabilities of such a god. Self-emulation that triggers an intelligence explosion becomes the catalyst.

Let us compare Emulationism with the idea of creation in typical religions. For instance, Christianity states that God created the world in six days [Carroll and Prickett, 2008]. Emulationism does not necessarily deny this but treats it as one possible scenario within the simulations used for creation. The idea that some entity has created a perceptual world is, when considered in light of the evolution of computers, improvements in simulation technology, and the advancement of AI, recognized as a phenomenon that could occur once human technology has progressed—akin to how we treat evolutionary processes in biology [Darwin, 1859] or the problem of global warming in earth science [Abbass et al., 2022]. However, as for creation taking six days or the details of how it was done, one cannot even label such claims as predictions or conjectures without an actual witness or credible, science-based proof. Yet if the cognitive world itself becomes emulable through a technological critical point, then what initially sounds like a wild notion becomes just one more plausible creation scenario. The same holds for evolution. Under Emulationism, both a scenario in which the cognitive world itself is generated and then life emerges and evolves, as well as a scenario in which a partially evolved world is created and continues to progress to the present, can be simultaneously accepted in a simulation-hypothesis framework.

The algorithm for creation, according to Emulationism, goes as follows:

1. *In a certain base reality, there exists an automaton that reaches the technological critical point.*
2. *That automaton generates within the base reality a perceptual matrix whose level of detail surpasses its own ability to distinguish (i.e., exceeds its discrimination threshold).*
3. *Repeat steps 1 and 2 in nested fashion.*

When steps 1 and 2 are repeated in a nested manner, new cognitive realities are continuously produced within the given "base" reality, one inside another. If the technological critical point is reached via digital information in a computer, then each newly created world within the base reality contains less discrete information than the base reality itself, and each sub-world in turn also contains proportionally less information. By the same reasoning, the process by which the base reality itself was created must also have emerged from a yet-earlier reality, but when we designate a certain standard as the base reality, all perceptual matrices—i.e., cognitive worlds—produced through nesting exist within that base reality. Consequently, we, along with superintelligences many "generations" before us in the nest, are all inhabitants of that same base reality. There is even a strong possibility that, like a scientist smiling next to a "brain in a vat," [Putnam, 2000] the owner of the computer that created our world is smiling right beside us. Of course, unless a superintelligence before us that generated the perceptual matrix has given us access (by definition), we have no way to perceive the "parent" world in that nesting. Nevertheless, once it has been demonstrated that reaching the technological critical point is possible, the scenario that our own world is a perceptual matrix becomes the most plausible theory. With intelligence explosions now being discussed as a realistically achievable technology, the notions of the technological singularity and the technological critical point are no longer mere speculation but phenomena we may well confront in reality.

4 von Neumann Probes and Fermi Paradox

Let us now discuss the Fermi Paradox [Armstrong and Sandberg, 2013] and how Emulationism addresses it in a consistent manner. Enrico Fermi is said to have reasoned that, given the age of the universe and the vast number of stars, if planets similar to Earth form with a certain probability around those stars, then extraterrestrial beings should be widespread throughout the universe—and some of them should already have arrived on Earth. His famous question, "Where is everybody?" captures the essence of the so-called Fermi Paradox [Herzfeld, 2019], which has been debated for many years.

Figure 3: Perceptual Matrix and Fermi Paradox

In the field of computer science, von Neumann proposed that an exponentially self-replicating probe [Matloff, 2022], sent to nearby planets, would be able to carry out large-scale mining operations on the Moon, the asteroid belt, and other planets in the most effective manner. Bostrom later argued that a powerful future superintelligence might realize such interstellar von Neumann probes. Tipler, on the other hand, asserted that because we have discovered no self-replicating spacecraft produced by a civilization other than our own, extraterrestrial civilizations do not exist—thereby reigniting discussion around the Fermi Paradox [Gray, 2015].

A variety of theories have been advanced to resolve this apparent contradiction: for instance, that extraterrestrial life is rare or nonexistent, that it goes extinct periodically due to natural phenomena, that intelligent aliens are insufficiently developed, or that the self-destruction of intelligent life follows from the natural order of things [Ball, 1973, Baxter, 2001, Bennett et al., 2016, Ward et al., 2000].

Emulationism provides an integrative explanation of these seemingly contradictory phenomena. First, Emulationism does not conflict with any of the above hypotheses. For each proposed scenario—whether we are truly alone in the cosmos, or whether some intelligent beings remain hidden from us—if a superintelligence exists that can run a perceptual matrix for that scenario, then the explanation "it is so because this cognitive world was created that way" holds true by definition of the perceptual matrix.

Figure 3 illustrates how cognitive worlds can be produced in nested fashion: starting with some baseline perceptual matrix (or "base reality"), then creating a new perceptual matrix within it, and so on, for n total nestings. If our cognitive world happens to be, say, the $n+1$th perceptual matrix, then there is a superintelligence at the nth level—along with its own cognitive world—and the intelligent entities we call "aliens" exist in that nth world. From their perspective, we are merely inhabitants of a simulation game; they can rewrite any scenario at will, branching possible timelines to produce whatever worlds they find convenient. Both we and they exist within the nth perceptual matrix, so we cannot verify what lies outside our own cognitive world unless they have configured it to be observable to us [Ishizaki and Sugiyama, 2024d]. Nevertheless, under Emulationism, we can say that advanced non-human intelligences exist in the same world—just not in our layer—and thus the ostensibly paradoxical question raised by Fermi can be resolved without contradiction. Moreover, this hypothesis becomes empirically testable once humanity triggers an intelligence explosion and reaches the technological singularity.

As noted in the Appendix A, with the development of language model technologies, humanity is closer than ever to realizing automata capable of technical self-emulation. If an intelligence explosion—and the accompanying perceptual matrix—are achieved, Emulationism will soon be empirically validated alongside the emergence of a superintelligence.

5 The Beginning and the End of Nesting

In the creation algorithm of Emulationism, as depicted in Figure 3, once it becomes possible to simulate a cognitive world, does the chain of nested worlds have a beginning or an end? To consider this question, we need to understand the nature of a world simulated within a computer. In a simulation, environmental information is represented as a series of discrete scene data. In a perceptual matrix—a

cognitive world—what we call "time" is merely one "frame" among many, and the moment we are experiencing is just one "page" of this information. If the simulation is represented as a set of discrete data in the form of matrices on a computer, then there must be a finite amount of memory available to store that data.

When a perceptual matrix is used to generate yet another perceptual matrix in nested fashion, as shown in Figure 3, let M_n denote the total memory of the nth perceptual matrix. Although M_n is large enough to create the nth perceptual matrix, the memory of the $n+1$th perceptual matrix must be less than or equal to M_n minus the constant C required to construct the simulator for the next perceptual matrix. Thus, for all natural numbers

$$\forall n \in \mathbb{N}; M_n > M_{n+1} + C \tag{1}$$

and because the complexity of the automaton must exceed the discrimination ability D in order for it to qualify as a "world" (i.e., to be cognitively recognized as reality rather than an obvious simulation), it follows that

$$\forall n \in \mathbb{N}; M_n > D \tag{2}$$

Let $M = M_1$. By induction from (1), for all $n \geq 2$,

$$M_n < M - (n-1)C \tag{3}$$

Since M is a constant, (3) shows that a necessary condition for M_n to satisfy (2) is, for all $n \geq 2$,

$$n < 1 + \frac{M - D}{C}. \tag{4}$$

From (4), we see that there is a limit to how many nested layers of worlds can exist starting from the base reality, indicating that there is a sort of "end of the world" at the small-scale limit. If a simulation is created whose level of detail can be discerned by the automaton's discrimination ability D—akin to a modern VR game—then cognitively speaking, it would not be called a world or a reality. It would only be "one simulation within this reality."

What about the parent-nested worlds that simulate our own base reality? Under Emulationism, there must exist a cognitive world with enough memory to emulate our reality. By induction, it follows that each such parent-nested world successively exists, each possessing a finite but greater amount of memory, implying that a perceptual matrix with an even larger memory might exist. From a purely logical perspective, one could say there is an infinite regress of parent worlds, so while there is no identifiable "beginning," there is a definable "end."

6 Conclusion

Emulationism generalizes the process by which new perceptual realities arise within our existing reality, through humanity's triggering of an intelligence explosion, reaching a technological critical point, and generating an over-real. Via the nested structure of successively created perceptual matrices, we not only offer a philosophical explanation for the Fermi Paradox—interpreting the existence of extraterrestrials from the perspectives of self-emulation and imitations of reality—but also reconcile evolutionary theory and creationism within an Einsteinian deterministic worldview, free of contradictions. If a perceptual matrix is realized, humanity may find itself lost in a continuously nested simulation worldview, one in which the beginning remains unobservable while only an end can be definitively identified.

Appendix A LLM Emulation

Efforts to reproduce intelligence using deterministic finite automata trace back to von Neumann [1966]'s self-reproducing automata—akin to living organisms—and Turing [1950]'s discussions of intelligence and emulation, subsequently developing under Wiener [1961]'s founding of cybernetics. Rosenblatt [1958] created the Perceptron, an early neural network that laid the groundwork for neural computation and learning processes. This was further enhanced by methods such as backpropagation [Rumelhart et al., 1986], enabling the learning of increasingly complex and deep structures. These groundbreaking

inventions allowed machine learning models, particularly in the field of natural language processing (NLP), to handle progressively more sophisticated information processing tasks.

With the introduction of Attention and the Transformer architecture by Vaswani et al. [2017], it became possible for networks to retain and represent more complex contextual information, leading to dramatic performance improvements in various tasks such as question answering, text classification, and image captioning. Subsequently, a series of pre-trained language models like BERT and XLNet were released [Devlin et al., 2019, Yang et al., 2019], substantially accelerating progress in AI-driven language tasks. Building on this trend, the few-shot learning paradigm inspired the emergence of large language models (LLMs) that learn more complex contextual information under a unified architecture, using massive parameters [Brown et al., 2020].

LLMs have begun to surpass human scores on a range of benchmarks—exhibiting capabilities in logical reasoning, mathematics, and even programming that were previously unattainable by smaller language models. Consequently, research has commenced on realizing self-emulation: generating LLMs using LLMs within a given programming environment. Leike and Sutskever [2023] have articulated plans to use AI in pursuit of AI researchers and superintelligence, aiming to develop "super-alignment" techniques to keep autonomous LLMs in check without direct human oversight [Ishizaki and Sugiyama, 2024e]. Sutskever et al. [2024] and [Altman, 2024] have declared that superintelligence lies within the realm of possibility, and are likewise working toward safe superintelligence and alignment.

Self-emulation by LLMs is not a concern exclusive to OpenAI. Ongoing research explores automating the full cycle of generating, running, and debugging LLM training programs through recursive self-improvement (RSI), leveraging techniques such as DPO and Self-Play [Rafailov et al., 2023, Chen et al., 2024]. Observing that LLMs continue to follow scaling laws and enhance their optimization abilities even beyond parameter counts that exceed human cognitive capacity, it is clear that using massive computational resources to train increasingly intricate models could lead to AI self-reproduction, potentially sparking an intelligence explosion—the "expandable approach" invoked by Chalmers [2010]. Ethical discussions surrounding this possibility have already begun [Puthumanaillam et al., 2024].

As living organisms evolve and improve their various goal-achievement capabilities, Morris [2003] contends that convergence is the dominant force in evolution; given similar environmental and physical constraints, life inevitably evolves toward optimized body plans, and at some point, evolution will arrive at intelligence—akin to what we currently observe in mammals. Transformer-based language models, exemplified by the Attention mechanism, are thought to contain certain crucial functional parameter sets—akin to "knowledge neurons" in the human brain's information processing [Dai et al., 2022]. If it is possible to emulate these essential functionalities, then such models possess an extraordinary scalability that views human-level intelligence merely as a stepping stone.

References

K. Abbass, M.Z. Qasim, H. Song, M. Murshed, H. Mahmood, and I. Younis. A review of the global climate change impacts, adaptation, and sustainable mitigation measures. *Environmental Science and Pollution Research*, 2022.

S. Altman. The intelligence age, 2024. URL https://ia.samaltman.com/.

D.J. Amit and N. Brunel. Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. *Cerebral Cortex*, 1997.

S. Armstrong and A. Sandberg. Eternity in six hours: intergalactic spreading of intelligent life and sharpening the fermi paradox. *Acta Astronautica*, 2013.

J.A. Ball. The zoo hypothesis. *Icarus*, 1973.

J. Baudrillard. *Simulacra and Simulation.* University of Michigan, 1994.

S. Baxter. The planetarium hypothesis - a resolution of the fermi paradox. *Journal of the British Interplanetary Society*, 2001.

C.H. Bennett, R. Hanson, and C.J. Riedel. That is not dead which eternal lie: The aestivation hypothesis for resolving fermi's paradox. *Journal of the British Interplanetary Society*, 2016.

N. Bostrom. Are you living in a computer simulation? *Philosophical Quarterly*, 2001.

T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020.

R. Carroll and S. Prickett. *The Bible: Authorized King James Version.* Oxford University Press, 2008.

D.J. Chalmers. The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 2010.

Z. Chen, Y. Deng, H. Yuan, K. Ji, and Q. Gu. Self-play fine-tuning converts weak language models to strong language models. *he 41st International Conference on Machine Learning (ICML 2024)*, 2024.

D. Dai, L. Dong, Y. Hao, Z. Sui, B. Chang, and F. Wei. Knowledge neurons in pretrained transformers. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

C. Darwin. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* John Murray, 1859.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

A. Einstein. On the electrodynamics of moving bodies. *Annalen Physics.*, 1905a.

A. Einstein. *Relativity: The Special and General Theory.* Springer, 1905b.

R.H. Gray. Von neumann probes: rationale, propulsion, interstellar transfer timing. *Final publication appeared in Astrobiology*, 2015.

N. Herzfeld. Where is everybody? fermi's paradox, evolution, and sin. *Theology and Science*, 2019.

R. Ishizaki and M. Sugiyama. Large language models: Assessment for singularity. *philpapers.org/rec/ISHLLM*, 2024a.

R. Ishizaki and M. Sugiyama. The age of superintelligence: capitalism to broken communism . *philpapers.org/rec/ISHTAO*, 2024b.

R. Ishizaki and M. Sugiyama. Ultimate intelligence and ethics. *philpapers.org/rec/ISHUIA*, 2024c.

R. Ishizaki and M. Sugiyama. Rsi-llm: Humans create a world for ai. *philpapers.org/rec/ISHRHC*, 2024d.

R. Ishizaki and M. Sugiyama. Self-adversarial surveillance for superalignment. *philpapers.org/rec/ISHSSF-2*, 2024e.

W. James. Are we automata? *Mind*, 1879.

C.G. Langton. Self-reproduction in cellular automata. *Physica D: Nonlinear Phenomena*, 1984.

S. Legg and M. Hutter. A collection of definitions of intelligence. *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms—Proceedings of the AGI Workshop 2006*, 2007.

Jan Leike and Ilya Sutskever. Introducing superalignment, 2023. URL https://openai.com/index/introducing-superalignment.

G.L. Matloff. Von neumann probes: rationale, propulsion, interstellar transfer timing. *International Journal of Astrobiology*, 2022.

H. Minkowski. Nachrichten von der gesellschaft der wissenschaften zu göttingen. *Mathematisch-Physikalische Klasse*, 1908.

S.C. Morris. *Life's Solution: Inevitable Humans in a Lonely Universe*. Cambridge University Press, 2003.

L. Muehlhauser and A. Salamon. Intelligence explosion: Evidence and import. *Singularity Hypotheses: A Scientific and Philosophical Assessment.*, 2012.

J. Neumann. Theory of self-reproducing automata. *University of Illinois Press*, 1966.

D. Peterson and M. Silberstein. Relativity of simultaneity and eternalism: In defense of the block universe. *Space, Time, and Spacetime*, 2010.

V. Petkov. Is there an alternative to the block universe view? *Philosophy and Foundations of Physics*, 2006.

K.R. Popper. Science as falsification. *Conjectures and Refutations*, 1963.

G. Priest. What is philosophy? *Philosophy*, 2006.

G. Puthumanaillam, M. Vora, P. Thangeda, and M. Ornik. A moral imperative: The need for continual superalignment of large language models. *arXiv:2403.14683 [cs.CY]*, 2024.

H. Putnam. *Brains in a Vat*. Oxford University Press, 2000.

R. Rafailov, A. Sharma, E. Mitchell, C.D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems 37 (NeurIPS 2023)*, 2023.

F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 1958.

D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning representations by back-propagating errors. *Nature*, 1986.

K. Schilbrack. The concept of religion. *Stanford Encyclopedia of Philosophy*, 2022.

B. Schöne, J. Kisker, L. Lange, T. Gruber, S. Sylvester, and R. Osinsky. The reality of virtual reality. *Frontiers in Psychology*, 2023.

C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948.

I. Sutskever, D. Gross, and D. Levy. Superintelligence is within reach., 2024. URL https://ssi.inc.

A. Turing. Computing machinery and intelligence. *Mind*, 1950.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems.*, 2017.

L. Wachowski and L. Wachowski. The matrix. *Warner Bros.*, 1999.

X.J. Wang. Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 2002.

P.D. Ward, D. Brownlee, and L. Krauss. Rare earth: Why complex life is uncommon in the universe. *Physics Today*, 2000.

D.J. Watts and S.H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 1998.

N. Wiener. *Cybernetics*. The MIT Press, 1961.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R.R. Salakhutdinov, and Q.V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems 32*, 2019.