Anton V Sinitskiy

¹Affiliation not available

December 22, 2023
New Restrictions on AI from Physics: The Most Reliable Way to Predict AGI future?

Anton V. Sinitskiy

ML LC, Massachusetts, USA
Email: anton.sinitskiy@gmail.com

Abstract

In recent years, advancements in Artificial Intelligence (AI) have accelerated, edging us closer to achieving Artificial General Intelligence (AGI). However, alongside these developments, there has been a surge in public concern regarding potential unpredictability and adverse effects of AGI. We hypothesize that the laws of physics, particularly those of Nonequilibrium Thermodynamics, may introduce unexplored constraints on AI. These constraints could be constructively leveraged to predict, monitor and limit potential future paths in the development of AGI. The proposed physics-based approach to understanding AGI includes viewing AGI as a part of a bigger system (the real world), emphasizing the significance of computing costs, establishing links to biological phenomena and applying mathematical tools for rigorous analysis. We hope that this framework will promote an innovative direction in research, adding another powerful method to the toolkit for understanding possible trajectories of AI, alleviating public apprehension and shaping the future discourse of AI research.

Introduction

The field of Artificial Intelligence (AI) and Machine Learning (ML) is rapidly evolving, with recent advancements marking remarkable progress toward the realization of Artificial General Intelligence (AGI). This accelerated growth stems from significant achievements in diverse areas such as natural language processing, image recognition, autonomous driving and complex game-playing, all of which hint at the potential for AGI in the near future.

As we move closer to this technological frontier, the complexity of the challenges posed by AI also rises, becoming as multifaceted as the intelligence we aim to develop. These escalating concerns surrounding AGI primarily revolve around the alignment of AGI goals with human values and interests, its potential for uncontrolled recursive self-improvement and an inherent lack of transparency and accountability, often referred to as the “black box” problem of AGI. Concerns about weaponization of AGI, the concentration of power in the hands of a few tech giants and the significant potential for economic displacement exacerbate the apprehensions. These issues are not just abstract ethical considerations, but are grounded in real-world implications that could impact societies at large in the near future.

In response to these challenges, there have been increasing appeals for regulations to govern the development and use of AGI. However, in practice, control may remain limited due to a multitude of factors. For instance, the geopolitical landscape can hinder the implementation of consistent global regulations, as governments may have differing perspectives and interests in the advancement and control of AGI. This conundrum represents a delicate balancing act between fostering innovation and ensuring ethical, safe AGI development.

Under these circumstances, the worst-case scenario might be an uncontrollable and potentially catastrophic development of AGI. However, AGI, like all systems, will be subject to laws that stem from its inherent workings and the physical reality within which it operates. These laws could form the boundaries that ultimately constrain the evolution and behavior of AGI systems.

In this work, we hypothesize that physical laws, particularly the laws of Nonequilibrium Thermodynamics, may impose as-yet undiscovered restrictions on AGI. We posit that these constraints, once understood and formalized, could offer a tangible framework to predict, monitor and limit possible future trajectories of AGI development. Such insights could form a foundation for developing robust, universally applicable guidelines for AGI, enabling us to steer the progress of AGI in a direction that benefits humanity.

Known Physics-Based Constraints

To clarify our intention to uncover universal constraints applicable irrespective of a specific material implementation of AI, we briefly discuss two physics-based constraints of this sort that are already known.
One fundamental limit on AI (including AGI is Landauer’s principle, which is derived from the laws of thermodynamics. It states that there is a minimum energy cost, namely $kT \ln 2$, associated with the irreversible erasure of one bit of information, where $k$ is the Boltzmann constant and $T$ is the absolute temperature at which the operation is performed. The efficiencies of currently existing computers and nervous systems are far from this theoretical limit.

Another example of a physical limit on AI is the Bekenstein bound. It constrains the maximum amount of information that can be stored within a given finite region of space with a finite amount of energy by a value proportional to the surface area of that volume, rather than the volume itself. The Bekenstein bound is applicable to information storage in both the natural and artificial realms, including potential AGI systems.

In the current state of computing devices, it appears that our current technology is far from saturating these two physical limits. However, as we delve deeper into miniaturization and densification of information, this gap may decrease over the years. Besides that, such limitations may be important from the conceptual viewpoint. For example, in cryptography, fundamental limits on the complexity of decryption play an important role in the practical choice of encryption algorithms.

Formulating new limits, similar to Landauer’s principle and the Bekenstein bound in their universal character, but preferably more restrictive under current technologies, could enable more accurate long-term predictions about the development and implications of AGI, and give us a clearer roadmap of this unprecedented technological territory.

**Principles of Search for New Constraints**

As we venture into the unknown landscape of potential constraints on AGI, it is important to lay out a few guiding principles that, in our opinion, should frame the investigation. None of these principles taken separately are new, but we believe that their combination may result in novel insights.

Firstly, we emphasize the necessity of perceiving AGI as an intrinsic part of the larger system (in principle, of the world as a whole), rather than an isolated abstract entity. At the most fundamental level, AGI systems are inseparable from the rest of the world and exist within the same unified system governed by the same laws. Subsequently, in the language of Theoretical Physics, this unity undergoes “a spontaneous symmetry breaking”, partitioning it into an intelligent part (the AGI) and the rest of the world. From this viewpoint, the perceived characteristics of AGI are in fact a consequence of the known properties and laws of the whole system to which AGI inherently belongs.

Next, we propose an increased focus on the cost of computing as a fundamental factor. While AI’s inner workings may seem abstract, the algorithms themselves are implemented in physical systems and their output interact with the physical world. As computing power is a limited and often expensive resource, the energy and time requirements of computations present natural constraints on large-scale AI (in particular, AGI). Loosely speaking, you can run any code – even a very inefficient one – on your personal computer, but as soon as it goes up to the scale of the recent large language models, let alone the estimated scales of AGI, deep optimization of efficiency becomes unavoidable. This perspective takes us on a different theoretical path than the numerous universality theorems for various types of ML models, which have often neglected practical considerations of the cost of computations. By taking into account the physical limitations of computing, we may discover new theoretical avenues that reflect the realities of AGI development.

In addition, we advocate for a closer conceptual connection between biology and AI. Insights from biological systems, which have evolved over billions of years, could offer valuable lessons for AI development. As some biological systems are already examples of intelligent systems optimized to operate within the constraints of physical laws, understanding these systems may lead to a better understanding of the potential constraints of AGI. This could be a fertile ground for bridging the gap between natural intelligence and AI. This analysis should focus not only on the similarities, but also on the differences between artificial and biological systems, making analogies between biological evolution and AGI development more justified.

Lastly, we posit that a rigorous, quantitative mathematical theory would be more beneficial than verbal or qualitative analysis in exploring potential constraints. Mathematical formulations can provide specific, quantifiable predictions that can be tested and refined over time. A quantitative approach is particularly crucial in dealing with complex systems like AGI, where verbal or qualitative intuitions may fail due to the high-dimensionality and nonlinearity of the system. By placing mathematical rigor at the core of our inquiry, we aspire to make tangible progress in identifying the boundary conditions of AGI.

**Simplest Model**

To clarify the principles discussed, let us take a look at a simple model we have recently introduced. This model takes inspiration from a speculative scenario regarding the emergence of the nervous system. It hypothesizes that primitive nervous systems might have evolved in Dickinsonia (or maybe other Proarticulata) during the Ediacaran period. These organisms fed on bacterial mats through external digestion and occasionally moved to new feeding sites. Accidental overlap of one organism atop another could harm the one below due to this external...
digestion process. Therefore, evading behavior of such organisms relative to each other should have been strongly supported by natural selection. Inherent traits of Proarticulata, such as cellular sensitivity to external electric fields and coordinated movement via cilia, might have acted as evolutionary pre-adaptations to the initial nervous system. While our model uses this scenario as a reference, its main mathematical framework and conclusions can be applied more broadly than just this specific origin scenario for the nervous system. We intentionally kept our model as simple as possible. In particular, it comprises only two dynamic variables: one representing the nervous system state (may be interpreted as the membrane potential of the neuron closest to a predator or an aggregated level of nervous system arousal) and the other indicating the state of the environment (the distance to the nearest predator).

The key results from the model include the following.\textsuperscript{47,48,50} Evolutionary optimization of the primitive nervous system tends to produce sharp, spike-like neural activations rather than gradual potential changes. After optimization, the nervous system is as highly sensitive to predators as possible when they are nearby, but this does not extend to predators at a distance. The optimum solution links the nervous system to effectors, enabling quick movement responses to potential threats. Functioning costs of the nervous system can be minimized with faster neural activation and higher sensory sensitivity. Evolutionary optimization also reduces noise in neural responses, minimizing false positive or false negative events (no escape in the presence of a predator or escape in the absence of danger).

While these outcomes might appear self-evident given what we empirically know about biology on Earth, it is crucial to highlight that they are strictly derived from broader principles (evolutionary optimization based on expressions from Nonequilibrium Thermodynamics). This gives us confidence in potential future research aimed at identifying general neural network properties that are not bound by their material makeup. Our methodology, based on the latest advances in Nonequilibrium Statistical Physics, is not confined to a specific form of dynamic equations. Even now, our approach offers intriguing insights, but we believe that expanding it to encompass larger neural networks will yield even greater insights. Future research might delve into larger models, admitting functional specialization into sensory, motor and inter-neurons.

**Lessons from the Simplest Model**

The primary goal of a neural network may be formalized strictly in mathematical terms, possibly using Nonequilibrium Thermodynamics. Our model is premised on maximizing the fitness of a simplified biological agent with a neural system. While fitness is a biological concept and cannot be directly applied to artificial neural networks, a growing understanding is linking fitness to Nonequilibrium Thermodynamics concepts, such as the growth or reproduction rates, entropy production rate, energy flows, dynamic stability or resilience and other related quantities.\textsuperscript{51-61} Contrarily, previous literature on fundamental theory of neural networks assumed models minimized other functionals, for example, root-mean-square error of predictions.\textsuperscript{52-66} We expect that fitness and its generalizations within Nonequilibrium Thermodynamics will provide a more solid basis for defining an AGI goal function than those previously proposed. By recognizing AGI as a part of the real world, we can examine it as adapted to long-term existence in the real world, rather than just learning the real world from aside. Comparative analysis of biological vs. artificial neural network goal functions will allow for improved extrapolations from biological phenomena to AGI. We hope further research in this field will shed more light on possible AGI goals, including their strict mathematical formalization, which would address one of the main above-
mentioned concerns regarding AGI. We do not claim here that fitness (or its thermodynamics-based generalization) will automatically arise as the goal of AGI, because AGI does not necessarily emerge through natural selection. However, we believe that such analysis may provide a better toolkit for formalization, providing at least building blocks (such as the potential function $u$ in our model) for the formulas for AGI goals.

A neural network model may initially treat neuronal and environmental variables holistically before separating them into two distinct classes. The simple model mentioned above is two-dimensional, with one variable relating to the environment and the other to the neural network state, allowing for the equal treatment of both.\textsuperscript{47,48} This approach reveals a complex connection between these variables, leading to non-trivial predictions (for example, a sharp response of the neuronal variable to sensory information as a consequence of a certain dependence of the fitness on the environmental variable). Unlike some previous models that focused solely on the neural network, our model integrates in an untrivial way the laws governing both the external world and the neural network dynamics. Pursuing this research direction, we hope to address concerns of uncontrolled self-improvement of AGI, yielding AGI progress predictions comparable (at least, in principle) to predictions about the physical world.

A mathematical, quantitative approach could bring to light significant details that verbal analysis may miss. Our model provides examples of such details, including border solutions (with infinite parameters) for the optimal dynamical equations that were not evident with a verbal analysis, the identification of missing constraints on the solenoidal components of these expressions and, on the other hand, an arbitrary contribution to the solenoidal component not determined by the potential, and, finally, qualitatively wrong consequences of the widespread weak noise assumption.\textsuperscript{31,44} This precision contrasts with many prior AGI studies that drew conclusions solely from verbal reasoning. We believe that mathematical rigor should significantly improve all aspects of AGI analysis and future predictions.

A general yet instrumental theory could possibly apply both to artificial and biological neural networks. Our model was inspired by a potential reconstruction of early nervous systems,\textsuperscript{49} but the analysis was conducted at a more general level, eschewing specific references to molecular or biological mechanisms. Such an approach to neural networks could improve our understanding of biological and artificial network similarities and dissimilarities\textsuperscript{67} and guide us more effectively in extrapolating from well-studied biological networks to forthcoming AGI systems. This could help to address public concerns about AGI more effectively by drawing parallels to familiar biological phenomena.

Open Questions

Further advancements in delineating the physical constraints on AGI may be achievable by addressing the following queries.

Firstly, we must identify a generalized form to express the cost of neural networks, both artificial and biological, without any need to reference specific material manifestations of these networks. While our current cost expressions are more broadly applicable than those in earlier literature, they are not yet universally applicable. We believe Nonequilibrium Thermodynamics offers a promising avenue for addressing this issue.\textsuperscript{50} By interpreting the entropy production rate as the cost of a dissipative process, successful attempts have been made in the literature to apply this perspective to several biological systems.\textsuperscript{51} However, this approach is yet to be used with a nervous system. Extending this to a neural network is not straightforward, as we have previously demonstrated.\textsuperscript{50} A mere minimization of the entropy production rate as the sole component of the cost results in a null value of the solenoidal component. Our hypothesis is that the entropy production rate can indeed be interpreted as the cost of the neural network, while the aforementioned mathematical anomalies may be resolved, perhaps, through dimensionality reduction from a higher-dimensional initial system representations or additional conditions on the solenoidal component necessary for the existence of the stationary distribution. Encouraged by the success of our model, we are hopeful that generalizing the cost of neural networks will not prove overly complex and that the resultant expression will keep its mathematical productivity.

Secondly, we must determine the comprehensive set of principles necessary for an ab initio derivation of optimal neural network properties. Our previous work\textsuperscript{47,48,50} proposed several such principles, including the interpretation of system dynamics in terms of potential function and solenoidal component, the utilization of evolutionary optimization with fitness (or its generalization) as a goal functional, and adherence to conditions governing the properties of potential function and solenoidal field during optimization which are not modifiable by the evolutionary process (physical laws of the external world). Is this set of principles exhaustive? Unlike, as we could not derive a general expression for solenoidal component from these principles alone.\textsuperscript{48,50} Are there other conditions imposed on physically accessible potential functions and solenoidal fields? Are there other cost components to consider? Or do the minimal descriptions of a neural network necessitate a higher-dimension model using these principles? These questions require further exploration.

We posit that answers to these questions will underpin future research on the predictability of AGI as neural networks constrained by their working costs. Although premature to make extensive forecasts, we illustrate it by several conjectures for future confirmation or refutation. These conjectures are shaped by
apparent similarities between possible AGI development and biological or societal evolution. Further investigation of common and contrasting features of the costs and goal functions of AGI vs. biological and social systems will either substantiate or invalidate these comparisons. For example, AGI might fracture into multiple autonomous entities, much like human society that has never unified into a single state. AGI might become an essential component of the biosphere, with growth bounded by specific material resources. AGI might establish internal resource allocation mechanisms, analogous to financial systems but grounded in physical resource quantities. Such development would make the macro-level progress of AGI predictable through quantitative tools akin to those currently used in macroeconomics.

AGI may evolve complex "social structures" reminiscent of those seen in animal societies, forming alliances, specializing roles or competing for resources. AGI might develop symbiotic relationships with humans or technological and biological systems, offering computational services or problem-solving capabilities in exchange for energy, connectivity or other resources. A coevolutionary dynamic between AGI and their environment, inclusive of human society and technology, might emerge, instigating an ongoing cycle of reciprocal influence and adaptation. These scenarios extend far beyond the reach of current quantitative methodologies. Nevertheless, there is a strong demand for the development of such mathematical methods. We believe the principles that we outlined above and exemplified by our simple model will play a crucial role in deriving such solutions.

Conclusion

In this study, we have proposed four key principles which will hopefully provide a new perspective on the physical constraints of AGI and potential insights into its future development. Firstly, we emphasize the principle of viewing AGI as a this-worldly entity. By perceiving AGI as an active participant in the physical world, we can build a theory on the known natural laws, better assess AGI adaptability and predict its progress, thereby mitigating concerns of uncontrolled self-improvement. Secondly, we acknowledge the integral role of computing costs. The energy dissipated by AGI networks during operation aligns with Nonequilibrium Thermodynamics principles, offering significant insights into the constraints on the AGI structure, dynamics and potential for self-improvement. Thirdly, we advocate for the exploration of connections to biology. Drawing lessons from biological neural networks can improve extrapolations to AGI, and by analyzing costs and goal functions in AGI compared to biological systems, we can gain deeper insights into AGI development. At the same time, with a systematic, fundamental approach, we can avoid making superficial, unjustified analogies between biological systems and AI. Lastly, we encourage the use of quantitative tools for analysis. By applying mathematical rigor, we can unveil significant details otherwise overlooked in verbal analysis, enhancing our understanding of AGI and making it less biased or prejudiced. These principles collectively inform our ongoing approach to delineating a generalized form for the cost of neural networks and identifying principles necessary for derivation of optimal neural network properties. Although this field of study remains exploratory, we believe these principles will underpin future research on AGI predictability and development, offering a new perspective on AGI and its role within the real world.

References


(38) Verdecchia, R.; Sallou, J.; Cruz, L. A systematic review of Green AI. *WIREs Data Mining and Knowledge Discovery 2023*, *13* (4), e1507.


