Author: Josh Giesbrecht

Chapter One

No Singularity: On the Mathematical Impossibility of an Unchecked AGI Explosion

Introduction

In this paper, we argue that a "Singularity" of infinitely-self-improving artificial general intelligence is not a rational prediction to make. The oversimplified argument for an imminent Singularity of digital learning rests heavily on the assumption that learning rate has no fundamental limits, and learning goals are possible to achieve in times approaching zero. By looking more closely at the process of learning itself, we can see that learning is limited by the rate at which one can affect the system to be learned. Furthermore, we will describe how all practical learning done by artificial intelligence also includes human interactions in the designing, preparing and high-level evaluation of learned

outputs, which also presents a hard limit on learning time reduction.

~~~~~~~

The promise of Artificial General Intelligence largely lies in the dreams of a Singularity, so named as a nod to the mathematical notion of a graph approaching infinity. In this case, the graph in question would be a function of "intelligence", and the singularity would be a steep rise to infinity. This extreme notion is held to be a logical conclusion of the following argument:

1. An artificial intelligence could be built with sufficient capability to improve upon itself.
2. Once an AGI can improve itself, it will then be capable of learning how to improve itself even more.
3. As this repeats, the rate of learning and thus the level of intelligence will approach infinity.

And so, an exponential gain is imagined, based on the idea of a proportional gain in ability at every step.

It should be noted here that the very idea of a general "intelligence" measure is hardly backed by either computer
~~~~~~~

science or psychology. Furthermore, an emphasis on IQ and other simplistic measures of intellect track closely with the eugenics-aligned aspects of the TESCREAL worldview.

For the sake of the counterargument presented here, we use time required to learn rather than IQ or some other general intelligence measure. When considering a set learning goal, we can then consider the rate of learning to be inversely correlated to the time required to learn. This allows us to map the Singularitarian concept of an increasing singularity onto rate of learning rather than the poorly-defined "intelligence". It also lets us reframe the overall problem as whether or not time required to learn for a given learning goal can approach zero. If the time required to learn cannot approach zero, the overall rate of learning cannot continue to rise exponentially and a 'Singularity' becomes impossible.

Initial Analysis

A general learning process can be modeled by a simple feedback loop as follows:

<loop diagram wheee>

The time taken to achieve a given external learning goal is a straightforward summation of all the iterations of this cycle taken until the desired model is achieved.

Time to goal = Sum i to n [Oi + Ei + Mi]

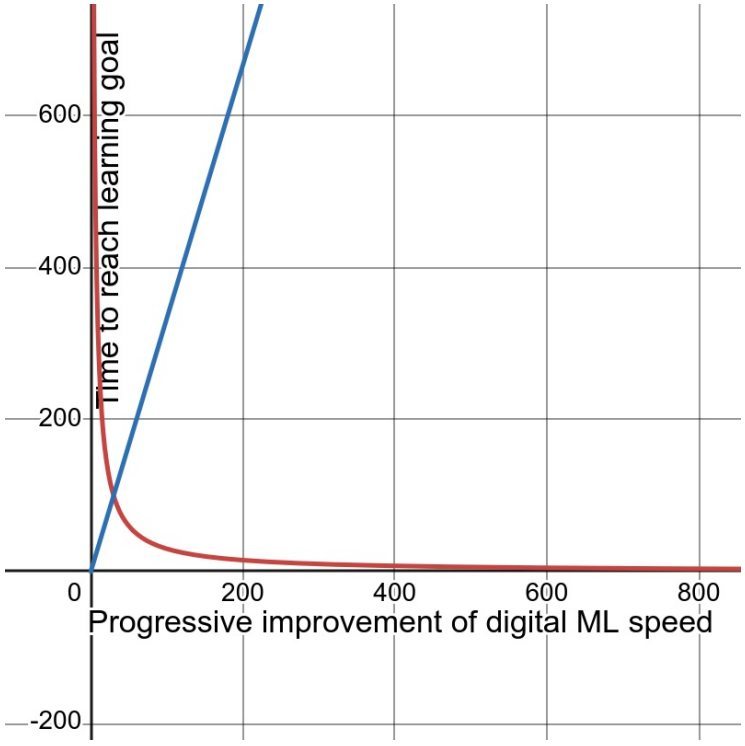
n = number of iterations required

Oi = output generation time

Ei = evaluation time

Mi = model revision time

For a learning 'singularity' to occur, all three of these terms would need to reduce to near-zero as i approaches n.



In the case of human learning, output generation could be anything from the writing of an essay draft, completion of a final exam, or crafting and delivering a presentation for a live audience.

Evaluation time can vary; in education, a distinction is made between "formative" and "summative" assessment. Formative assessment, or "assessment as learning", generally has shorter feedback loops that can impact the methods and tactics of learning by directly informing the student and/or the instructor involved before the process is complete. This could include "check your understanding" questions given to students with an answer key, or one-on-one dialog with an instructor about the topic being learned. As such, formative assessment leads naturally to "model revision time", or learned improvement on one's understanding.

Summative assessment refers to the sort of reporting-out that happens at the completion of a course, such as a final exam grade. It can be seen as an 'output' for external consideration intended for not only the learner and teacher but also parents, future school application bodies, or administrators and school boards. We can also think of this as a longer-duration feedback loop on the learning process itself. A learner who ends a course with a 45% will need to reconsider their approach to learning on a retake; a teacher who has many students failing a course might end up in discussions with external parties involved.

Artificial learning can also be described through this model. In the next sections we will give some basic examples from common deep learning strategies.

~~~~~~~

Deep learning is defined as machine learning tasks where the learning takes place in layered neural network models. The term 'deep learning' began when models may have had only five to ten layers of artificial 'neurons', though today's models can easily have a layer count in the hundreds.

A deep learning classifier is a model which is trained to apply specific categorical labels to the input, based on training from pre-labeled data. A classic example of this would be models trained to recognize handwritten digits from data such as the NIST image set.[1]

Let us apply the learning model above to this ML procedure. For each iteration, the output term Oi would be the time taken to process one set of inputs through the existing neural network and generate an output. The evaluation term Ei would be the comparison of the generated output to the training data's classification for that input. The model-

1. Ref needed here
~~~~~~~

adjustment Mi would be a backpropagation pass to adjust the model's weights to (hopefully) improve the next iteration's output and close the gap between desired and generated output.

Notably, the number of iterations required for a typical deep learning classifier is many orders of magnitude higher than, say, the formative assessment feedback loops of a high school student studying Algebra. A human student might do tens of questions a day to train, write ten or twenty quizzes, and a half dozen or so unit tests. A deep learning model will likely go through hundreds of thousands of individual iterations, grouped into epochs and processed in parallel batches to expedite the process.

In the above description, all three terms will reduce as applied computing speed increases. Output generation, evaluation, and model adjustment are limited only by the processing hardware's capabilities. Does this mean that classifiers can form the basis of an effective learning 'singularity'? In practice, it seems not. Classifiers are only as effective as the range of their training data and the accuracy and precision of the labels. Classifiers are unable to learn anything from unlabeled data, and sometimes training data can 'teach' the model the wrong pattern by accident.

While there is much to be said on how to improve on this process and avoid the problems mentioned above, there is a more fundamental point at play here that we will highlight shortly. Before we do, though, let us look at another model which finds a way to bypass many of these issues entirely.

~~~~~~~

Over the past decade, generative deep learning paradigms such as the Generative Adversarial Network, or GAN, have exploded in popularity. In generative models, the desired output is no longer simple labels, but rather a synthetic product that is hard to distinguish from the original source material set while still being unique (to some degree). This can take the form of generated images, or synthetic text generation through transformer-based models such as ChatGPT, Claude, or Grok.

GAN models took a page from the strategy book of game-playing AI models who were trained in pairs to competitively improve themselves quickly. Rather than both 'players' taking similar actions, though, a GAN pairs up a media-generating model with a fake-or-real classifier model. Both "generator" and "discriminator" progressively improve their game, with the generator learning to make better 'fakes' and the
~~~~~~~

discriminator improving its ability to spot them. Inputs to a GAN can vary; the most basic example uses an abtract input vector to give the model a mathematical framework to build patterns onto. More advanced GAN models take an input prompt meant to align in a human-meaningfully way to the output. A popular example of this are the image generation models that take a text prompt, such as Stable Diffusion.

While GANs are performing a different sort of task than a classifier, creating media rather than categorizing it, we can see how this approach enables advanced learning without (necessarily) requiring prelabeled data. This opens up the possibility of learning from gathered data without the extensive human step of filtering and labeling it ahead of time.

Today, In 2024, the most invested-in and hyped AI models are Large Language Models, or LLMs. LLMs are structured slightly differently than a GAN, but follow a similar paradigm of imitative generation. LLMs are trained on a massive data set harvested from public internet web sites, digitized books and journals, and so on. By using the large-scale data set of publications and the public internet, proponents have hoped to build models that can effectively learn from a vast corpus of

human knowledge to build an artificial intelligence that can approach the hoped-for 'singularity'.

However, what is a LLM actually learning how to do? Similarly to a GAN model, an LLM attempts to create output which is indistinguishable from source data output, which in this case is human-written digital text. The model generates output based on what kind of answers it has seen for similar input in its training data.

At no point in this process would either an LLM or GAN be able to "score" higher by creating an output that is superior to human output. In fact, in the case of a GAN, this would make it easier for the discriminator to spot the synthetic output, and so the generator would learn to avoid such a thing! The pinnacle of either model of training is to produce something effectively indistinguishable from its source data set - unique, but blending in perfectly. As such, generative models are quite literally incapable of ever learning something superior to the corpus of data it is fed.

Corporate proponents and makers of LLMs are seeking to build off of their foundations and solve this limitation by using further input as a form of Reinforcement Learning. In

one article by ____ and ____, CEOs of (OpenAI and whoever, ugh), Reinforcement Learning (RL) is held up as a sufficient learning model for all intelligence, on the argument that organic life and human intelligence has developed solely as a form of long-term RL over generations. And so, while we are not told openly exactly how these models are continuing to learn, we can safely assume that ChatGPT, MSWhoever, and etc are using human responses to their generated texts to try and optimize and further improve future output.

Reinforcement Learning, however, highlights a key problem which has been lingering in this entire Machine Learning endeavor for those who wish to build a 'singularity'. In this next section we will look at this problem in more detail, and reconstruct a more accurate model of a general learning process.

An Improved Analysis

In comparing these digital models to the simple learning model described for human learners, we might see a parallel in the matter of "formative" and "summative" assessments and how they serve different purposes. For the most part, the Ei evaluation term we have described has arguably been formative, directly tied into immediate feedback for the "learner" in as short a feedback loop as possible.

Is there an equivalent "summative" assessment step in training AI? We can see the end product of this kind of assessment in nearly every generative AI research paper in the form of example output. This assessment demonstrates that learning is relevant and accurate. It is also reasonable to assume that any researcher working in the field would also be

reviewing such output themselves during the process. After all, left to their own devices AI solutions have a known tendency to easily "solve" the wrong problem. These false solutions come in the form of overfitting to the data set[2], finding categorical patterns in data that do not actually match the desired real-world pattern[3], or by generating output which seems correct to the competing discriminator but has structural problems obvious to a human viewer[4].

No matter what we choose to call it, these human interactions are currently an existing necessity in an artificial learning process. In fact, we can usually find human interaction baked into the early stages of the process as well. Human coders choose what data sources to scrape for a training data set; potential inputs are categorized and labeled; generated outputs are reviewed, rated, and potential problems or offensive results flagged.

To accurately analyze the speed of the entire learning process of artificial intelligence, then, we must incorporate

---

2. Add a reference to something here (You Look Like A Thing maybe)

3. And here

4. And here

human interaction as well.

~~~~~~~

Let us look at a more complete and precise way to describe the time used in the learning process.

As described before, the process of training a generative model has a tight feedback loop where the model learns to synthesize output that "blends in" with the training data as closely as possible. We model this algebraically as

Time to goal = Sum i to n [Oi + Ei + Mi]

However, to situate this within the entire learning process, we must recognize this as an inner loop which is contained within a larger feedback loop guided by human engineers, designers and end users.

Time to goal = Ph + Sum j to n [Sum i to n [Oi +Ei + Mi] + Eh + Mh]

$$\sum_{j}^{m}\left(Eh+\sum_{i}^{n}(Ei+Mi+Oi)+Mh+P\right)$$

Ph = human-driven preparation
~~~~~~~

Eh = human-driven evaluation of output:

Mh = human-driven revision of the learned model

Human-driven preparation (Ph) includes selecting, gathering and formatting a data set, coding the initial conditions and network architecture, and setting the measurable learning goal used by the digital process.

Human-driven evaluation (Eh) includes testing output for over-fitting, degenerate-case pattern matching, and other situations where AI models get stuck on unhelpful "solutions" or local maxima.

Human-driven model revision (Mh) includes revising the training data, altering the network architecture, or designing filters for generated output to avoid offensive

results[5]. It can also include modifying the internal "goal" that is set by the digital process so that it more closely aligns with the real-world desired goal. (More on that later.)

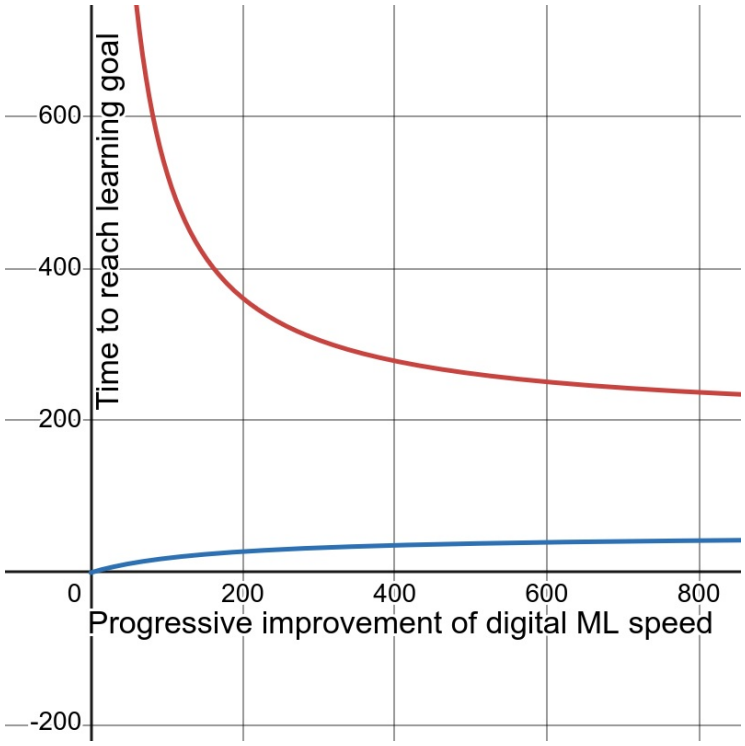
Now we can see a serious problem with our hopes to reduce learning time down to near-zero. While increasing our computation speed and capacity can (arguably) reduce the purely digital learning times, they have no direct effect at all on the human-driven terms. The increase in capability caused by massive-scale LLM engineering enables huge amounts of work to be done in the inner loop, but does not (on its own) eliminate the time needed for human design, supervision, and moderation.

The change in results can be seen graphically below. While the shape of the overall graph looks similar for "Time to reach learning goal" (in red), it notably has a horizontal asymptote well above zero. The impact on this on the so-called "intelligence" measure (in blue) is clear. Instead of linearly progressing ever upwards, it slows and plateaus rapidly.

---

5. Reference for this? Or just mention how much of a pain it is that none of the major corps actually reveal how they band-aid this BS

Simply put, with the constant terms of necessary human interaction involved, there is no singularity--only a reached, maximum limit of this so-called "intelligence".



Approaches to Mitigation

To overcome this obstacle, a Singularitarian must plausibly demonstrate that human interaction can be either be reduced so as to be inconsequential, or replaced by automation or further AI. We will look at each of these possibilities in more detail so that we can identify exactly what each claim requires.

~~~~~~~

Our first possibility for analysis would be to attempt to minimize the impact of human intervention, without removing it completely. To do so, we would need one or both of the following changes:
~~~~~~~

- decreasing the number of human-interaction iterations required (j)
- parallelizing the human-interaction iterations

The other obvious possibility is reducing the time humans take per iteration. However, I would posit that this is simply not going to have significant impact. Improving workflows might shave time taken down somewhat, but even a drastic improvement such as cutting time per interaction in half is still many factors short of the scale of time reduction needed to resemble a singularity-like learning improvement.

Decreasing the number of human interactions required sounds great, but what would that entail? It would be easy to argue that fewer mistakes will be made by future AI, thus less oversight required, but this is a circular argument. The interactions are currently required; we cannot dismiss them outright by saying, "A superior AGI will not require oversight" if the learning process cannot become a superior AGI while they are required.

A slightly more coherent argument would be to aim for a more generalized learning goal, and we can see this in the

current wave of LLM development. Earlier language models could predict the desired characters, words, or a few sentences in response to an input. Today, we have AI models that can be requested to generate much larger textual outputs, ranging from a few paragraphs to entire articles. By training on a wider range of textual data, the LLMs are also able to work towards "successful" outputs on a wide range of topics at once.

By casting a wider net, can we simply churn our way to a superior AGI? Many Singularitarians seem to believe so, or even claim that today's LLMs have already done so. And yet we regularly find that LLM output should not, in fact, be left unsupervised. (example example example lololol)

~~~~~~~

To understand why this is unavoidable, we need to make a distinction between the quantifiable goal which a digital learning process can iterate towards and the social goal which the designer of the process is hoping to achieve.

A designer of an LLM project might tell themselves, "Our goal is to train this model on the sum of human intellect as represented on the Internet." However, the actual, computable
~~~~~~~

task of these models is to be *difficult to distinguish* from the pool of source data. The entire nature of the process is to mimic what they are given, and potentially give us interesting patterns "in between" the example data. A GAN or LLM process is never rewarded for generating output that is *better* than the source data--how would the system know? An artificial neural network can only train on what it can quantify, and these learning models have no way to measure "better" or "worse", only "different". Technically, "better" would actually be penalized by the discriminator in a GAN (or trained away from in an LLM) and the model would quickly unlearn that behavior.

What this means is that while "time to learn to mimic internet text" might be something the system can be optimized for, "time to learn how to create superior responses than the usual internet reply" cannot be, with anything resembling our machine learning models.

In general, any time the digital system is optimizing for a quantifiable goal which deviates from the desired real-world design goal, we must include real-world evaluation (Eh) and correction (Mh) to the whole process. Since AI is only able to optimize towards quantifiable goals, and many goals in life

are not easily quantifiable, this presents a serious limitation.

~~~~~~~

One way to circumvent this is to parallelize the human interaction component, in an attempt to widely distribute the extra time spent and thus speed up the overall completion. After all, if one thousand people can simultaneously do the real-world evaluation and adjustment, that will take 1/1000th of the time. We have seen this approach widely used by companies such as Google, who have used their Captcha 'security' measure to label massive amounts of computer vision data via human interaction.

Reinforcement learning (RL) is a machine learning architecture that attempts to place the entire evaluation process outside of the learning agent and into the hands of a third-party evaluator. RL creates a system by which learning agents are given a 'reward' for good results, and it is left up to the agent to try different strategies to optimize for rewards. To some who hope to build an advanced AGI, RL is seen as a guaranteed way forward, since arguably the search for
~~~~~~~

'reward' is how intelligent life evolved in the first place[6].

Could RL with widespread "volunteer" human evaluators lead to an advanced AGI intelligence? Two significant problems quickly become apparent.

First, crowdsourcing the evaluation efforts would necessarily mean that the learning tasks would have to be creating outputs that everyday people would recognize as valid solutions or not. This limits the problems to ones which the average populace understands well enough to identify valid results. It also limits the possibility of superior solutions being recognized as such, for example if they go against the expectations of the human evaluator. Far more likely is that in any complex situation, the AI model would simply learn the expected "wisdom" of the crowd. A possible exception to this would be solutions which were unexpected but could then be tested for validity - but more on that problem in the next section.

Second, this approach to crowdsourced evaluation does not

---

6. That “Reinforcement Learning is Enough” propaganda er I mean article by those OpenAI / Google honchos

afford a true "singularity" in learning, as we are still limited by the time taken to distribute results, have them evaluated, and process those human evaluations into our digital system. We may reduce our learning time by a factor of a thousand, million, or even billions, but at some point we have a limited amount of volunteer person-hours available to do the RL process's evaluation work for us.

~~~~~~~

Finally, let us look at the holy grail of AGI singularity proponents - the possibility of simply removing human limits from the entire process. Why not create AI learning processes which take in data from the outside world and then evaluate solutions themselves? What if we create high-level "reward" models to solve complex but quantifiable and measurable problems such as mortality rates, poverty levels, or CO2 emissions? If a system can measure these results directly, no human interaction need limit the speed of learning.

For smaller systems, this is perfectly tractable. A small learning model can be designed to optimize for specific measured results in an external system - a thermostat that optimizes for reaching specific temperatures at a given time, for example, without having to be programmed with a precise formula of how to do so.
~~~~~~~

What would this look like in a larger system? Can we simply design a huge deep learning model that reads in a (literally) global variable and tries to optimize it? What would its outputs be?

First, and obviously, an artificial system trying to optimize a large-scale, real-world output must have the freedom to directly impact the factors affecting that output. The learning thermostat controls when the heat goes on - how would the learning Climate Change Manager control emissions? The question of what actions such a system could safely take control of is incredibly significant. Would the keen Singularitarian CEO hand over city planning, global food production, military power, etc to a Deep Learning toddler and let this AI fumble around with potentially lethal consequences? What disasters would take place before it even achieved parity with human expertise?

There is also the question of what it would be maximizing for. For example, could our baby Singularity solve climate change by simply shutting down industrial production? If not, how does it measure industrial stability along with environmental goals? Whose industrial stability would be

prioritized?

We could go on, but we should very quickly realize that while quantifiable measures like global temperature, GDP, and mortality rates can be good indicators of humanity's problems, simply choosing which of these to optimize for and within what parameters is prone to all sorts of human biases. And isn't guiding and selecting these measures to evaluate on exactly what was already happening in Eh and Mh terms for our previous systems? Isn't human intervention still required, then?

Finally, even if we somehow pick the ideal "global variables" for our infant AGI to optimize towards, we still have one unavoidable problem: time. None of these processes happen quickly. The learning thermometer can afford to take a few days or even weeks to perfect its process, but more importantly, it has no other choice but to do so. The learning feedback loop of Output -> Evaluate -> Model adjustment is taking place in a real-world environment where evaluation cannot take place until the changes set out by the output have had time to propagate and affect the outside world. Heat generated in a furnace has to spread through the ventilation system across the whole building before a measurement elsewhere can update and evaluate its success. Every physical

and social system we live within follows the same simple property.

Output -> physical system is affected -> Evaluate

This inherently limits the rate of learning. As such, we can propose the following: all learning about the physical world (or any external system) is limited in speed by the rate at which changes to the system propagate and affect the whole.

We have an irreducible constant term in our learning rate - the need to interact with reality. No amount of digital acceleration can erase it.

~~~~~~~

A common dream within Singularitarian stories is that of simulated worlds. These myths have a long history, much older than digital computation, in the form of philosophical and theological speculation such as Descartes' dreaming mind-in-a-jar hypothesis. Perhaps this is how our Singularity avoids the slow pace of experimenting and testing theories in sluggish reality - could it not learn from a simulated world at a much faster pace?
~~~~~~~

We dismiss this as a solution to the problem for the simple reason that it is begging the question. One cannot simply bypass the logical limits of the road to a Singularity by staking claim in the solutions found at the destination.

Modern simulations present a useful tool to make and study predictions, critique various models of systems, and yes, to give learning agents a virtual space to work within. But we need to be incredibly clear that AI agents learning within a simulation are learning *about the simulated world*, not reality itself. Existing simulations of physical or social systems are many orders of magnitude less complex than reality itself. We can teach a robotic arm to find efficient routes to a physical problem with a simplified simulated physical world, but we are nowhere near being able to simulate the complexities of a whole body's biological chemistry, a brain's full network of signals, or the detailed social intricacies of global politics and economics.

Many of the most celebrated examples in AI research of unsupervised learning are in games such as Chess or Go, which have relatively large possibility spaces and, culturally, an aura of intellectual achievement. However, no matter how

complex these games are, they can be learned in a highly accelerated manner by large AI systems because they are clearly and completely self-contained rule-systems with no necessary connection to real-world limitations. In other words, AI systems are free to learn gameplay at an accelerated pace precisely because the rules of the game world itself requires no interaction with the physical world.

Learning from games and simulations is not going to give us a world-solving Singularity of self-improved learning. Any transfer of learning from game-world to reality is going to require the same real-world testing described above, meaning the constant terms of human and/or physical world interaction.

Closing thoughts and conclusion

In this paper we have focused on the strict notion of a near-infinite Singularity, with learning times continuously approaching zero. One can imagine a simple argument - why not accept continued improvement even if it does have a constant non-zero limit? Could we not still reap the benefits of such a near-Singularity anyway?

We certainly can -- but we should do so with a realistic outlook, and that first and foremost requires throwing away the pervasive TESCREAL mindset in upper-management AI development that hinges on Singularity-type results. This means no more hand-waving away legitimate criticism on the engineering limits of machine learning, no more ignoring the

costs to the environment and to human rights with promises of magic futures, and in general recognizing that any benefits from ML systems have a limit and a real cost.

With machine learning seen as a potentially useful but limited tool, we should also remember to consider other avenues of problem-solving. Using our best learning hardware, applied in a widely parallelized learning architecture, with increasingly improved learning algorithms and strategies, is a goal worth working towards. However, the best digital learning systems pale in comparison to the complexity of the human mind. This author would argue that this goal is more likely to be accomplished through Universal Basic Income and well-funded public education.

~~~~~~~

The oversimplified argument for an imminent Singularity of digital learning rests heavily on the assumption that learning rate has no fundamental limits, and learning goals are possible to achieve in times approaching zero. By looking closely at the process of learning itself, we have seen that learning is necessarily an interaction with a larger system to be learned about, and thus the learning itself is limited by the rate at which changes propagate throughout the system to
~~~~~~~

be understood. Furthermore, we have seen that all practical learning done by artificial intelligence also includes human interactions in the designing, preparing and high-level evaluation of learned outputs, and thus is limited overall by the time taken for said human interaction. With both of these considerations in mind, it is clear that the time taken to achieve a real-world learning goal can never approach zero, as these limitations add a constant amount of time which digital improvements cannot simply remove.

In short, a "Singularity" of infinitely-self-improving artificial general intelligence is not a reasonable prediction to make, and there is no logical path towards building such a thing that can succeed.

# # # # #