On The Impossibility of an AI Mathematician Being Both Autonomous and Useful

August 11, 2025

Kharen Musaelian
Qognitive, Inc.
kharen@qognitive.io

Abstract

We consider a slight modification of the Birch test for an AI mathematician. We replace the original requirement that the AI satisfy the conditions of producing mathematical work that is IAN (interpretable, autonomous, and non-trivial) with a more succinct formulation, AU (autonomous and useful). We demonstrate that no AI in the form of a Turing machine can pass the modified Birch test.

1 Introduction

Recently, Y-H. He and M. Burtzev proposed the Birch test [1] as a benchmark for AI's ability to do mathematics autonomously. The three requirements are that the AI produce work in mathematics that is interpretable, autonomous, and non-trivial (IAN). It is the combination of the three that is hard to achieve. For example, some of the recent achievements of application of AI in mathematics are clearly non-autonomous [2]. It is also very easy to create autonomous AI that would apply rules of deduction to produce an infinite number of new "theorems", however all of them trivial. An example would be extending the multiplication table to ever larger numbers *ad infinitum*. It has already been demonstrated by Turing [3] that Hilbert's decidability problem, namely whether all mathematical truths within a well-defined class can be resolved by a Turing machine, can be restated as the halting problem, and the answer to the latter is 'no'. The question we are concerned with is whether a fully autonomous Turing machine (ATM) can discover any useful math at all. We use a construct similar to Turing's halting problem [3] to demonstrate that an ATM cannot pass the Birch test. Although inspired by and similar to Penrose's proof [4] that a human

mathematician is not a Turing machine, the present proof is distinct from it, while contributing to Penrose's overall argument.

For the purposes of our derivation, we will slightly modify IAN into AU, replacing the combination of interpretable and non-trivial by "useful". Clearly, the words "interpretable", "autonomous", and "non-trivial" are not sufficiently well defined. However, substituting "useful" for "interpretable and non-trivial" does not change the substance of the definition, while allowing it more precision, as we will demonstrate below. Our goal is to show that under any reasonable definition of the words "autonomous" and "useful", as applied to mathematical work produced by a robot, the two conditions cannot be satisfied at the same time.

2 Demonstration

Consider the infinite but countable set of all possible Turing machines $\{R_k(n)\}_{k\in\mathbb{N}}$ that produce mathematical work given an input natural number n. We will posit that these RMs produce all sorts of intermediate results, and only upon finding a useful result, may stop and print out the entire derivation. This means that unless a useful result is attained, the machine will keep running and deriving more and more intermediate results, which it will not deem "print-worthy". The condition that an RM halt only upon attaining a useful result is essential, because otherwise the machine cannot be **autonomous**, i.e. a human would have to check the printed results for "usefulness". One example of an RM that should not halt is the above mentioned multiplication machine, as it produces a lot of correct mathematical work that is not useful (trivial). It is important that the converse is not true, i.e. the machine upon finding a "useful" result, is not **required** to halt. We can think ATM as an orchestrator that algorithmically launches RMs as bottom-up theorem factories that simply apply rules of deduction to the body of previously known results and, upon completing each instruction, run a specialized Turing machine called Detection Machine (DM) that detects if the execution so far has produced "useful" mathematical work. Note, that since there is no requirement to **determine** that the work is useful, we can make sure that DM will always halt. We can simply put a limit on its number of operations, at which point DM gives up without ascertaining that the work is "useful". The existence of such a DM is not obvious, but it is a requirement, since the RM has to detect the "usefulness" autonomously. We have not so far defined what constitutes "mathematical work", but in order for ATM to spawn RMs algorithmically, the set of RMs needs to be recursively enumerable. Hence, we posit **axiomatically** that "mathematical work" can be defined in a way that RMs are recursively enumerable [1].

Now, the problem is that an RM will almost certainly be stuck in an infinite run, never producing a useful result. Therefore, the Birch test can only be passed if there is a Turing machine, which we will call anticipation machine

[1] The author is indebted to Dimitrios Tsementzis for pointing this out

(AM), that can determine if an RM $R_k(n)$ will halt. This would be analogous to the human insight that arguably precedes all our derivation steps [5].

Note that AM, to be usable, must always halt. However, if such an AM exists, then there also exists an Extended Anticipation Machine (EAM) that, upon determining that a given $R_k(n)$ halts, puts itself into an infinite loop.

Note: While an RM is a Turing machine, not every Turing machine is an RM. Here we make an **axiomatic** stipulation, that EAM, i.e. a Turing machine which upon halting demonstrates that $R_k(n)$ does not halt, has attained a result that is "useful". This is the only condition on "usefulness" that we will have. It is also an entirely reasonable and natural stipulation, as under any reasonable definition of "usefulness" that some of the $R_k(n)$ could satisfy, it has to be useful also to know if some $R_k(n)$ will not halt!

As the reader would have noticed, we formulated the problem of EAM in terms of the Turing halting problem. Below we follow a variant of the proof of the Turing halting problem [4] to show that AM does not exist.

Lemma. There is no RM that can determine if $R_k(n)$ will halt for all k and n.

Proof. Assume that EAM exists. Namely, $A(k, n)$ will denote the EAM that halts if $R_k(n)$ does not halt, and does not halt if $R_k(n)$ halts. Now consider $A(n, n)$. It is an RM, as it is a Turing machine that will only halt upon attaining a useful result. Hence it is on the list of R_k, and thus $A(n, n) = R_g(n)$ for some number g. Now consider $R_g(g)$. By construction, if $R_g(g)$ halts, it shows that $R_g(g)$ does not halt! Hence, $R_g(g)$ does not halt. But the only way it does not halt is for the original AM to have determined that $R_g(g)$ halts! Therefore, by *reductio ad absurdum*, the Anticipation Machine does not exist.

Theorem. RM cannot be both "autonomous" and "useful".

Proof. Given that Anticipation Machine does not exist, it is not possible to predict algorithmically that $R_k(n)$ will not halt. Therefore, any RM can end up in an infinite loop that will require a human intervention to reset.

3 Possible objections.

Objection 1. There is only one example of R_g for which $A(k, n)$ fails. Why couldn't an RM be useful most of the times? It doesn't need to be always useful.

Response. The diagonal-slash proof is a form of *reductio ad absurdum*. It is like the Euclid's proof that there is an infinite number of primes. The proof does not show that there is "just one extra prime missing". Similarly, Cantor's original diagonal-slash proof that real numbers are not countable doesn't just discover one extra number. In fact, natural numbers constitute an infinitesimally small fraction of real numbers. Similarly, the number of non-stopping (hence "useless" RMs) is likely infinitely larger than the number of "useful" RMs.

Objection 2. Couldn't we design our RMs such that they will likely halt? Isn't this the whole art of creating AI?

Response. Here comes the issue of what is understood as "autonomous". One requirement for an RM to be autonomous is that humans don't have to sift

through reams of useless results, hence the machines should halt upon finding a useful result only. Another requirement is that the machines are not designed with a particular independently known result in mind. After all, the idea is that a machine does a somewhat brute force search and churns out results hereto unknown to humans. Indeed, human intervention can be thought of being between two extremes: 1) carefully sifting through drivel generated by an undiscriminating RM hoping to find, by sheer luck, a golden nugget; 2) using human insight and understanding to design computer programs to attain particular aims, for example [2] a computer program that calculates the next prime number. The former is hardly an appealing way of advancing mathematics, while the latter is commonplace, but can hardly be called "AI doing math".

Objection 3. In all of the above, the term "useful" is not defined. Wouldn't the argument apply to any other adjective describing a result obtained by an RM?

Response. While the term "useful" is not defined, the stipulation that $A(n, n)$ is itself an RM relies on a commonsense understanding of what it means to be useful. For example, if we replace "useful" with "green", there is no reason that a Turing machine that shows that an RM is "not green" is itself "green". Just as our argument is of the same flavor as Gödel's, Turing's, and Tarski's theorems, so is the concept of "usefulness" from the same category of concepts that allow self-reference as "provability", "computability", and "decidability".

Objection 4. Couldn't we avoid the problem of halting by setting some arbitrary cutoff on how long an RM can run before we halt it?

Response. It makes no difference. Since the goal is to produce a useful result eventually, upon halting a machine, we have to restart algorithmically another one, until some machine produces a useful result. Then, the entire chain of machines can be rolled into one Turing machine that only halts upon attaining a useful result.

Objection 5. What about the original Birch test?

Response. We replaced "interpretable and non-trivial" by "useful" to improve clarity. However, one could make the same argument with the original Birch test by stipulating that the result that a certain RM R_k given a certain input n does not halt is both interpretable and non-trivial. That should be just as uncontroversial.

Objection 6. Perhaps, "usefulness" of mathematical work can only be defined outside of mathematics.

Response. This is concordant with Tarski's theorem and we are sympathetic with that point of view. That would mean that DM does not exist. Of course, in that case the rest of the argument is unnecessary, as the absence of DM by itself proves that an autonomous RM cannot be useful.

[2] Due to Larry Moss

4 Discussion

The argument we present in this paper can be thought of as a formalization of the intuition that a computer program cannot determine autonomously if the results it produces are valuable or that "its line of thought" is fruitful. Formalization of fuzzy concepts, such as "interpretable", "autonomous", "useful", and "trivial", is bound to be somewhat controversial, and we have addressed some of the possible objections. However, we believe that the gist of the argument holds regardless of the particular choice of words, and has both philosophical and practical implications.

Firstly, our argument only applies to Turing machines. Indeed, unlike a Turing machine, an analog physical device may not have a countable number of states. This would suggest that human mathematicians develop useful mathematical work because they happen to be more like "analog devices", than Turing machines.

Secondly, our result casts doubt on exuberant expectations [6] that Turing machines can autonomously produce real mathematical work [3]. For better or worse, we will need humans to do math, even if armed with powerful computers.

5 Acknowledgments

The author would like to thank Luca Candelori for his indispensable help in sharpening the argument and Alexandre Abanov, Vahagn Kirakosyan, Larry Moss, Santhanam Nagarajan, Stefano Pasquali, Dimitris Tsementzis, and Dario Villani for useful discussions and valuable suggestions.

References

[1] Y.H. He, M. Burtsev, "Can AI make genuine theoretical discoveries?," *Nature*, vol. 625, 241, 2024.

[2] B. Romera-Paredes, M. Barekatain, A. Novikov, M. Balog, M. Kumar, E. Dupont, F. Ruiz, J. Ellenberg, P. Wang, O. Fawzi, P. Kohli, and A. Fawzi, "Mathematical discoveries from program search with large language models," *Nature*, vol. 625, 12 2023.

[3] A. M. Turing, "On computable numbers, with an application to the Entscheidungsproblem.," *Proc. Lond. Math. Soc.*, 1937.

[4] R. Penrose, *Shadows of the Mind: A Search for the Missing Science of Consciousness*. USA: Oxford University Press, Inc., 1st ed., 1994.

[3] In the latter reference, mathematician jobs are reckoned to be some of the most likely to be replaced by AI, reflecting a view widespread among schoolchildren that math is the most robotic subject.

[5] I. McGilchrist, *The Matter with Things: Our Brains, Our Delusions and the Unmaking of the World. Volume I, the Ways to the Truth.* London: Perpectiva Press, 2021.

[6] K. Tomlinson, S. Jaffe, W. Wang, S. Counts, and S. Suri, "Working with AI: Measuring the occupational implications of generative AI," 2025.