doi: http://dx.doi.org/10.23925/1980-7651.2021v27;p01-12

How AI can be surprisingly dangerous for the philosophy of mathematics — and of science

Walter Carnielli *

Abstract

In addition to the obvious social and ethical risks, there are philosophical hazards behind artificial intelligence and machine learning. I try to raise here some critical points that might counteract some naive optimism, and warn against the possibility that synthetic intelligence may surreptitiously influence the agenda of science before we can realize it.

Keywords

Machine learning; Artificial Intelligence; Philosophy of science; Justification; Interpretability.

All these good things

Here [1] is a story. You have found a strange creature, perhaps an alien, who seems to have a tremendous mathematical capacity. She (or he, or it), the alien, starts talking about things you do not understand. If you do not give reasonable answers, maybe she is going to destroy you.

You think it strange that the creature cannot understand what you are saying, and has to send the words one after another. She seems to think this is funny and makes a smart remark about you as a person. You just try to ask some rational questions to check if the creature understands what you say. But she only makes another smart remark about you, which confuses you more and makes you angry.

Once again the creature shows that she does not understand what you are saying, and she only says: "Oh, you want another explanation, huh?"

You say "Yes, I need another explanation". She starts to make another explanation, but you interrupt her and tell that it is too simple. You explain that the first explanation was not enough, because it does not contain details.

You say "If this is a mathematical question, please give me the definitions!" The creature seems to understand you better and she starts to calculate with his hand.

You say "I see you calculating with your hands, but I do not understand what it means. Is this some mathematics from other civilizations?"

For some reason the creature seems to get angry with this question and starts to shout at you. She says she doesn't know what you are talking about. At this point you give up and stop asking her questions.

*ISTC-CNR Laboratory for Applied Ontology- Trento, Italy and Centre for Logic, Epistemology and the History of Science University of Campinas - Unicamp, Campinas, SP, Brazil

[1] This article was exceptionally produced using the LaTeX package to correctly arrange its content with mathematical formulas. Therefore, it presents a different format than the one commonly used by the Circumscribere Journal.

You seem to fear that the creature is thinking of destroying you, but maybe you are wrong. You say "Can't we be friends, and learn from each other?" The creature seems to understand what you are saying, but she says that there is no time.

You say "Please give me some example of the mathematics you know, that is new for people in the planet Earth". The creature sends a message through the eye-patch and you can read a lot of letters on it. You can't understand what she is saying, but suddenly you meet some new mathematical concepts that you never heard before.

This story was created by Generative Pre-trained Transformer 3, also referred to as GPT-3, through AI Dungeon [2], slightly edited to correct typos and to balance genders. AI Dungeon is a free text-based game built on top of GPT-3, which with human cooperation (adding questions) helps to build a story around any topic you choose.

The Generative Pre-trained Transformer 3 is a new text-generating neural network, whose name refers to a Google innovation of 2017 called 'Transformer' which can figure out the likelihood that a particular word will appear in connection with other words, with several new applications.

GPT-3 scares a lot of people, because it can mimic styles of literature, and even invent some fake quotes. It could be dangerous, because It could help to generate false text, like deepfakes, and helping to spread fake news. Created by OpenAI, GPT-3 is a product of the artificial intelligence research lab that was before sponsored by SpaceX and Tesla.

Artificial Intelligence (AI), also called synthetic intelligence, has advanced tremendously. Personal assistants like Siri, Google Now and Cortana are improving, and self-driving cars are already there. Super-personalized healthcare, applications in national security, new modes of transportation, new kinds of industry and business, and highly effective education, to name just a few, are promises for the near future. Statistical treatment of Big Data and advances in machine learning are certainly changing the world. But from the social and political viewpoint, there are some worrying aspects. AI can create mass unemployment, and widen even further the gap between those who have access to these technologies and the billions who don't.

Advances in AI can potentially destabilize the delicate balance that has prevented nuclear war since 1945. The physicist Stephen Hawking famously predicted that the emergence of artificial intelligence could be the 'worst event in the history of our civilization'.

The famous Turing Test, one of the seminal debates in the field opened by A. Turing in 1950,[3] can be regarded as a natural language processing problem: is it possible to build an AI that can convincingly pass itself off as a person? If anything can do it, GPT-3 is the best candidate up to now. If it is, we may have an example of real AI, but we will continue not knowing what intelligence is.

The quest for superintelligence

Thanks to the advances as the ones mentioned above, among others, there is a revival in the discussion on the potential calamitous fate of AI. Superintelligent AI stands amid the prospect of a disaster: an entity smarter than the best human brains in practically every field, as warned by Bostrom[4] , looks extremely dangerous.

A recurring theme in topics of AI security is the possibility of keeping a superintelligent agent in a sealed container, in order to prevent it from causing any harm to humanity. Bostrom devoted a significant part of his book to the problem of control, that is, to the problem of ensuring that a newly created superintelligence would act in accordance with human interests.

[2] https://play.aidungeon.io/.
[3] **turing1950**.
[4] **bostrom:2014**.

One of the solutions commonly proposed for the control problem is to restrict the access of superintelligent machines to the real world. D. Chalmers proposed in 2010[5] the idea of a 'leakproof' singularity (as the intelligence explosion is often referred to). He suggested that, for security reasons, AI systems should first be restricted to virtual worlds, until the behavioral tendencies of superintelligent machines could be fully understood under controlled conditions. It is like caging a monster for a while, until you understand its personality.

Containing AI can lead us to abandon its benefits. On the other hand, any form of communication with a contained superintelligent agent can be risky. Some people argue that any superintelligent AI should never be left outside the confinement cage, regardless of the circumstances.

There is still no absluteformal guarantee against superintelligence. We already live with things that seem smarter than the best human brains in restricted fields, a form of proto-superintelligence, and nobody knows when the limit of superintelligence could be reached. Today we run billions of computer programs on globally connected machines, without any formal guarantee of the absolute security of its consequences. It could happen, in principle, that a new application developed for smartphones would trigger a chain reaction that would destabilized the markets and the world stock exchange, or something much worse.

Could there be such an absolute guarantee against the potentially dangerous risks arising from superintelligence? This point is quite controversial, and it will be discussed in the last section.

In any case, the purpose of this article is not to praise the goodness of AI, nor to insist on its possible dangers. Not even to discuss what is, or is not intelligence (artificial or otherwise), or superintelligence. Long before any superintelligence - or even to be able to master it - we need to think with the philosophy that we have available. I wish to raise doubts and concerns about the future of science (and humanity) – it seems we are on the horns of a dilemma: either science succumbs to mere correlation rather than advancing into metaphysical models that include cause and effect, purpose, explanations, etc., or we teach machines to think metaphysically, and we all succumb.

While the second alternative still seems remote, the first is knocking at the door: synthetic intelligence seems to be already covertly influencing the agenda of science.

How mathematics underlies AI

AI and machine learning (ML) are obviously founded on usual mathematics — not easy, but relatively simple mathematics: to build clever applications you need Linear Algebra, multivariable differential and integral calculus, probability theory and statistics, complexity of algorithms and combinatorial optimization. To work on the more theoretical side you may need some topology, Fourier transforms, information theory and elementary logic. But you certainly do not need sophisticate topics on the foundations of mathematics, infinity categories, higher topos theory, algebraic geometry or homotopy theory — at least for now. This does not mean that someone will not invent a new ML paradigm based on chaos theory, or that the solution of the Riemann hypothesis will not lead to new trends in AI; but the fact is that AI and ML do everything they already do now based only on the relatively simple and well-known theories listed above.

However, AI and ML are able to do automated discovery in several areas. The generative modelling systems, for example, especially the generative adversarial networks (GANs), can repair images with missing pixels, or infer missing information by means of a competition after some adequate training data. Good examples are the generation of faces of people that never

[5] **chalmers2009singularity**.

existed. Well trained, GANs can build links between hair styles and the concepts of 'man' and 'woman' and quickly deduce a connection. Bayesian networks are another important tool in AI for representing probabilistic relationships between multiple events, able to make future predictions and to explaining (at least statistically) previously unrelated observations. The negative side is that some authors are so intoxicated by the apparent preponderance of probabilistic techniques that they foresee an explicit departure from logic towards statistical techniques in future AI[6].

But not all AI pioneers and people who made decisive contributions to progress of AI are overly optimistic. Judea Pearl, winner of the Turing Award 2011 and champion of Bayesian networks, thinks that "All the impressive achievements of deep learning amount to just curve fitting". Pearl claims that without causal reasoning machines will not reach human-level intelligence: "To build truly intelligent machines, teach them cause and effect", is the title of an interview of his for Quanta Magazine[7] .

In his book with Dana Mckenzie,[8] Pearl defends that the way to build really intelligent machines is to replace the current paradigm of reasoning by association with causal reasoning. Causal models would help robots to think counterfactually, and this would be a step to self-awareness and free will - which could of course be quite dangerous, unless we learn how to tame robots.

What is important here is that, even independently from these more recent advances in AI and ML, computers have helped mathematics to develop in the last four decades. Some simple but illuminating examples will be discussed next.

Computers helping to develop mathematics

Computers have definitely helped to solve certain types of mathematical questions that involve searching, or that involve proving exceptionally long or complex theorems. In principle, the task of finding a proof of a theorem is not a recursive task, but the activity of verifying the validity of a proof is by definition computable. It turns out, though, that it often exceeds human capacity.

Instead of proving it, we may be interested in finding a counter example, for instance in the domain of integers. Take as an example the famous Goldbach's conjecture, which states that every even natural number greater than 2 can be written as the sum of two prime numbers. There is a simple algorithm that permits one to test, for each even natural number $2n$, whether or not there are two primes p and q less than $2n$ such that $2n = p + q$. If we wish to falsilfy Goldbach's conjecture, we may keep trying this algorithm until we find a certain number $2k$ which is not the sum of two prime numbers. This kind of procedure is called recursively enumerable, computably enumerable, semidecidable, or Turing-recognizable[9] . The problem is that nobody has ever found such a counter-example, and this algorithm may run forever. Yet a computer can do the search a lot more quickly, and find a counterexample (if one exists) after billions of attempts in a fraction of the time that humans would take. Quantum computers, when available, even without changing the semi-decidability characteristic of the problem could even shorten this time logarithmically: they could potentially reduce the time it takes to run a search of this kind from billions of years to a few minutes. Some famous cases that have already benefited from standard computers are briefly discussed below.

[6] So is the case of E. Charniak, *apud* (**sep-artificial-intelligence**), who boldly proclaimed in a talk that 'logicist AI is moribund', and that the statistical approach is the only bet for the next 50 years.

[7] **quanta2018**.

[8] **pearl2018book**.

[9] **computability2018**.

The shortest paper ever

Leonhard Euler conjectured in 1769 that, for any integers n and k greater than 1 and $n < k$, ${x_1}^k + {x_2}^k + \cdots + {x_n}^k = y^k$ has no positive whole solution. (For $n = 2$ this is Fermat's last theorem, which we now know to be true). But in 1966 a counterexample was found by computational 'brute force'. Euler's conjecture was disproved by L. J. Lander and T. R. Parkin when, through a direct computer search on a CDC 6600, a counterexample was found for $n = 4$ and $k = 5$.

This was published in a paper comprising just two sentences, resulting in the shortest article ever published[10] . The counter-example is $27^5 + 84^5 + 110^5 + 133^5 = 144^5$, and possibly it would not have been found without the help of a computer. More counterexamples have been discovered since then, including for $n = 3$ and $k = 4$.

The reason why there are counter-examples for these cases, and not for $n = 2$ and $k > 3$ (corresponding to Fermat's last theorem) is a hard problem for number-theorists, one that computers helped to expose.

The Four Color Problem

The Four Color Problem was proposed in 1852 when Francis Guthrie, a young mathematician, was coloring a map showing the counties of England.[11] It occurred to him that the maximum number of colors required to color any map seemed to be four, if any two neighbor regions sharing a common boundary receive different colors. The resulting Four Color Problem became known as the second most famous open problem in mathematics, after Fermat's last theorem.

In 1976 Kenneth Appel and Wolfgang Haken[12] announced that they had solved the problem. But much of their proof was carried out on a computer, using 1,200 machine hours, equivalent to 50 days on the computers of that time. The proof was too long to be checked by humans.

Errors were found, which contributed to make many mathematicians unhappy that most of the proof consisted of brute force computation and that the essence of the proof could not be examined by humans. In each case, Haken and Appel quickly corrected the error.

Although we are still not completely sure if the Appel and Haken proof is correct, we are now certain that the Four Color Theorem is true because of another type of proof, also generated by computers. In 2008, George Gonthier and Benjamin Werner proved the Four Colour Theorem using the automatic proof-assistant Coq. A mathematical assistant is a relatively new kind of computer program that works in an interactive fashion, with a human providing ideas or questions in mathematical language while the computer carries out the verification or provides a counter-example.

This kind of assistant is itself certified to be absolutely correct when it works, even if we cannot prove the consistency of Coq by using Coq itself, as this would contradict Gödel's second incompleteness theorem[13] . But since we can prove the absolute correcteness of Coq, we can be sure that the automatically generated proof by Coq is correct (as much as mathematics itself is consistent). A mechanized proof of the Four Color Theorem based on the Coq proof assistant can be found at the GitHub repository [14].

The humanly non-verifiable proof of the Four Colour Theorem gave rise to a hot debate about the question of to what extent computer-assisted proofs count as philosophically acceptable proofs.

[10] **lander1966counterexample**.
[11] **wilson2013**.
[12] **appel-haken:1976**.
[13] **computability2018**.
[14] url=https://github.com/math-comp/fourcolor.

After Gonthier and Werner efforts of using a computer to check the work of another computer, though, the philosophical debate reaches another level.

The initial fear that the original program could have a hidden flaw which would fool everyone no longer makes sense; the second computer proof is based on a logical system, it implements a higher-order type theory, and in the case of the Four Colour Theorem, produced a surveyable proof. Mathematicians, logicians and philosophers should get used to the idea that from now on it will no longer be possible to restrict the verifiability of mathematical proofs to human scrutiny. The complexity of certain mathematical proofs seems to require new tools.

Kepler's cannonballs problem

In 1611 the astronomer Johannes Kepler conjectured that the maximum density of a sphere packing in a three dimensional space is achieved by the familiar cannonball arrangement, in the same way that grocers stack oranges.

Curiously, the seed of the idea came off the mind of Sir Walter Raleigh, explorer and 'official' pirate on the high sea, who committed his assistant, the English mathematician Thomas Harriot, to find a fast and reliable way to estimate the number of cannonballs in a munitions pile. Raleigh's interest was obviously military, or piracy. Harriot sought the help of Kepler, who was also interested in such stacking problems — interested, as he was, in relating the orbits of planets to the shape of perfect solids. Kepler eventually published a little booklet titled *The Six-Cornered Snowflake* (1611), which concerned the stacking and crystalline arrangements in ice, and ended up influencing the science of crystallography.

Kepler's conjecture turned out to be a problem that puzzled mathematicians for four centuries, one which was proved in 1998 by Thomas Hales and Samuel Ferguson. An essential part of the proof by Hales and Ferguson is a dense computer code, which cannot be completely verified by a standard human peer review process.

The first proof published by Hales and Ferguson in 1998 counted about 300 pages of text, and was based on 40,000 lines of computer code. The *Annals of Mathematics* refereeing process took 6 years, with referees declaring that they were "99% certain" of the correctness of Hales' proof.

To eliminate any uncertainties about the correctness of the proof, Hales launched the Flyspeck project in the beginning of 2003. The aim of this project was a complete formal verification of the Kepler conjecture using a combination of the Isabelle and HOL Light proof assistants.

In 2014, eleven years later, the Flyspeck project team, headed by Hales, announced the completion of a formal proof of the conjecture using such a combination of proof assistants.

Now, it's official: in 2017, the formal proof was accepted by the journal *Forum of Mathematics*, signed by no less than 22 authors[15] . The formal verification proof, just as the one for the Four Color Problem, is available at the GitHub repository[16].

This case provides further evidence of the growing movement of using computers to verify proofs, proofs that human beings produced with the help of computers but which they cannot themselves verify. What this suggests, and what some mathematicians already foresee, is that in the future computers will do all the heavy work, leaving mathematicians to ponder the deeper questions about what the proofs represent, and how they can be applied. Or, alternatively and more pessimistically, that computers will dictate the agenda of mathematical development, with unsettling consequences. We will return to this point later.

[15] **hales2017formal**.

[16] url=https://github.com/flyspeck/flyspeck.

Can computers influence the scientific agenda?

Computers have been used not only to verify, but also to build original evidence for some new problems.

Some years ago a computer system called HR conjectured several new number-theoretic conjectures which involved some novel concepts, conjectures that were proved by S. Colton.[17] They included some nice ideas like the following:

- If the sum of the divisors of an integer is prime, then the number of divisors will also be prime. For example, the divisors of 16 are 1, 2, 4, 8, 16, their sum is 31, so it has necessarily a prime number of divisors.
- The sum of the divisors of a square is always odd. For example, the divisors of 49 divisors are 1, 7, 49, with sum 57.
- HR also 'invented' the concept of refactorable numbers: a number is refactorable if the number of divisors is itself a divisor. For example, 9 is refactorable since 9 has three divisors (1, 3 and 9).

HR has conjectured several additional properties related to such concepts, which are true but not trivially imagined, or proved.

A much more surprising number-theoretical property was discovered by a man-computer team, the so-called 'Prime Conspiracy': the distribution of primes is not random! An intuitive assumption in number theory is that prime numbers behave like random numbers in a sort of normal distribution: any prime number must have an equal chance of being followed by one of the four terminations 1, 3, 7 or 9 (the four possible endings for all primes other than 2 and 5).

But this is not so. Kannan Soundararajan and Robert Lemke Oliver[18] , helped by a computer, analyzed the first 400 billion primes and found out that primes not only seem to avoid being followed by another prime with the same final digit, but also have 'preferences'. Primes ending in 3 seem to prefer to be followed by primes ending in 9 more than by primes ending 1 or 7. Primes ending in 3 or 7 are followed by a 1, 30 percent of the time instead of the expected 25 percent. Likewise, a prime ending in 9 is almost 65 percent more likely to be followed by a prime ending in 1 than by another prime ending in 9.

Why? Nobody really knows. Lemke says that primes 'hate to repeat themselves'. Be that as it may, this discovery, which arguably could never have been made by humans alone, naturally imposes new mathematical and philosophical directions. What other properties of primes, so fundamental, have we failed to observe? Could a team of computers working by themselves be able to propose conjectures and prove their own theorems?

One of the most important problems in the philosophy of science is the problem of induction. Even before the method was clearly formulated as part of the scientific method by the philosopher Francis Bacon in the 17th century, Sextus Empiricus, the Pyrrhonian skeptic of the 2nd century, had already questioned the validity of inductive reasoning, arguing that a universal law could not be established from an incomplete series of particular instances. But despite the millenia of criticisms, Bacon even influenced Isaac Newton in the discovery of his mathematical laws of physics from the data of the movement of the planets. A firm advocate of the method of generalizing from data, Bacon considered induction as the basis for how laws of nature are discovered.

[17] **colton2005automated**.
[18] **oliver2020distribution**.

Induction, it seemed, could be regarded as a mechanizable procedure to obtain scientific laws from data regularities: that is until another philosopher, David Hume, started to uncover its problems in the 18th century, somehow resuming the ideas of Sextus Empiricus and formulating his own skeptical arguments. It is impossible, sustains Hume in his well-known work, to establish any law by induction. It is always possible that there are black swans to revoke the 'law' (established by observation) 'All swans are white'. Actually this famously happened when Europeans arrived in Australia and saw black swans, and can happen to any law established inductively.

Karl Popper, in the 20th century, agreed with Hume in two notable ways. Popper struck two deadly blows against induction (or at least tried to). Firstly, Popper sustained that scientific discoveries are not mechanically induced from data. Instead, scientists confide in creative insights to propose new theories, and later conduce experiments to test those same theories. Popper discusses several aspects of induction, and states that induction "cannot be logically justified", thus taking logic out of the game;[19].[20] But Popper, in a joint work with D. Miller,[21] has also argued that there is no such thing as probabilistic inductive support. Their argument on the impossibility of inductive probability also takes probability, in special the Bayesian theory of evidence, out of the game.

There are several serious criticisms against the Popper-Miller attack, one of the most representative being the counterattack by I. J. Good[22] . It is hard to deny that probabilistic induction is possible, at least in some cases. Probabilistic theories are based on the presupposition that an increase in the empirical support of a conjecture increases the probability that the conjecture is true. For example, there is a consensus that smoking causes lung cancer, but this was in principle based on a correlation. Even if it is common knowledge that statistical correlations do not represent causal connections, probabilistic updates, applying for instance the Bayesian theory of evidence, may lead to new scientific laws. And this can be mechanized: machine learning algorithms start from huge amount of data and produce statistical models. Bayesian networks represent an important tool in connecting probability and causation. Induction, via machine learning, has challenged the barriers raised by Popper.

Partial Differential Equations are very useful for modelling air turbulence, weather patterns, and many other applications, but are notoriously hard to solve analytically. They are now being quickly solved using new deep-learning techniques.

In any event, now that AI increasingly invades the sciences, we need to be careful, even leaving aside important ethical issues, not to inadvertently modify the canons of the scientific method. AI and Machine Learning can be a gift, since they greatly shorten the time of discoveries, but they have a dangerous potential, because they can also replace science by hasty conclusions.

The case in mathematics

Paul Cohen, who won a Fields Medal for his work in mathematical logic, predicted that mathematicians would be replaced by computers 'at some unspecified future time'[23] .

Timothy Gowers, another Fields medalist of the University of Cambridge, thinks they may one day replace human reviewers at mathematical journals. This seems like a minor thing but it's not: reviewers help to dictate science and math agenda. Gowers believes that there is a good chance

[19] **popper2014conjectures**.
[20] **popper2005logic-sci**.
[21] **popper1983proof**.
[22] **good1990suspicious**.
[23] **future-of-proof**.

that human mathematicians may actually be out of the game by the end of this century. And once computers have enough capacity to prove theorems, they will also be able to decide which results to prove, liberating themselves from the need for human guidance.

This won't happen overnight, but it can slowly change the conception about the role of proofs in mathematics. For most mathematicians, proofs should explain why and how something is true, instead of merely establishing the truth. It is debatable whether computers can go so far as to produce explanations, and whether humans could ever understand such an explanation.

The frontier of what is considered to be 'artificially intelligent' moves constantly. Thirty years ago, we would consider very intelligent a program that translates instantly from almost any language to another. A program that would beat the best chess and go players in the world, or a candidate to pass the Turing test such as GPT-3 (see section 'All these good things'), would be seen as superbly intelligent. As we humans move the border, however, influenced by industrial, political and social tendencies, we are demanding more and more capacity from computers. Are we also getting ready to live with monsters?

The Russian mathematician Vladimir Voevodsky (Medal Fields, 2002) once discovered an error in one of his proofs, in a theorem that he considered quite relevant. This led him to propose a program for formalizing mathematics on the computer, launching a process that led to something more ambitious: a reformulation of the foundations of mathematics. The underlying intuition is that set theory is not the only way to do math: types can be used instead of sets, categorical thinking can replace and generalize the several ways of thinking mathematically. In this context, Voevodsky considers the proof of mathematics by computers as a necessity — this task has become too complex for humans. The world of higher mathematics is becoming too elaborate: Timothy Gowers also believes that the collaboration between a human mathematician and with a semi-intelligent database can 'take a lot out of the drudgery of research'.

Points in a geometric space behave like computer programs. The research program known as HoTT (Homotopy Type Theory) intends to illuminate the profound relationships between the theory of programming language, logic, and topology. A homotopy is a continuous deformation between two surfaces, and works as a way to classify surfaces that connects algebraic topology, logic, computer science, category theory and type theory (as pioneered by B. Russell, A. Church, and K. Gödel). In this way homotopy theory intends to offer a well-suited new foundation for mathematics that is at the same time amenable to computerized verification.

For more than a decade, Voevodsky defended proof assistants and developed univalent foundations to bring the languages of mathematics and computer software closer together. The Coq and Agda proof assistant programs, for example, are directly based on type theory. Voevodsky believed that human brains could not keep up with the increasing complexity of mathematics; he is reported to have answered the question whether all future mathematicians would end up using computers to create their proofs: "I can't see how else it will go."

Unfortunately Voevodsky passed away in 2017 at the age of 51. But his ideas are profound, and may in the future lead to 'mathematical silicon'[24] .

What do the Prime Conspiracy and Vladimir Voevodsky's proposal teach us? Regarding Prime Conspiracy, nobody (yet) knows: but this is a human conjecture, to be proved by humans, assisted by computers. This is very different from expecting computers alone to 'discover' or 'invent' conjectures which may have unknown relevance.

H. Putnam defends[25] that mathematics should be interpreted realistically (that is, objectively true or false, independently of the human mind). His paper emphasises the importance of quasi-

[24] **ladyman-presnell:2018**.
[25] **putnam1975**.

empirical, or even frankly empirical methods in mathematics.

To illustrate this point, Putnam appeals to a delicious fiction about Martian mathematics[26] :

> Let us now imagine that we have come in contact with an advanced civilization on the planet Mars. We succeed in learning the language of the Martians without too much difficulty, and we begin to read their newspapers, magazines, works of literature, scientific books and journals, etc. When we come to their mathematical literature, we are in for some surprises.
>
> What first surprises us is the profundity of the results they claim to have obtained. Many statements that our best mathematicians have tried without success to prove — e.g. that every map can be colored with four colors, that the zeroes of the Riemann zeta functions in the strip above the unit interval all lie on the line — appear as assertions in their mathematical textbooks. Eagerly we start reading these textbooks in order to learn the proofs of these marvelous results. Then comes our big surprise: The Martians rely on quasi-empirical methods in mathematics!

Should we be really surprised? If the Martians are so successful in using empirical methods, then why have we not used them? The fact, he says, is that we, earthlings, have been using quasi-empirical and even empirical methods in mathematics all along. Putnam gives as examples the quasi-empirically discovery of the correspondence between real numbers and points on the line. Another case is the discovery by Euler that the sum of the series $1/n^2$ is $\pi^2/6$. Putnam even suggests that model theoretic methods could be used to try to convert 'probability' arguments on number theory into proofs.

If Putnam is right, then computers might have a relevant role in providing quasi-empirical methods for the sake of mathematical discovery. But again, this is quite different from letting computers take the reins of science in their artificial hands.

The statistician George Box is famous for having said "All models are wrong, but some are useful." One of Google's research director amended George Box's lemma[27] revealing the hidden philosophy of companies, like Google, which have no commitment to science, and for which correlation supersedes causation: "All models are wrong, and increasingly you can succeed without them". For such economic giants, warns C. Anderson[28] , 'science can advance without coherent models, unified theories, or really any mechanistic explanation at all'.

Other authors, such as Napoletani, Panza and Struppa[29] , advert that the employment of big data science is already causing a significant change in the predictive power of mathematical analysis, modifying its role in the process of producing knowledge. The main point of big data analysis, they claim, is its disdain for causal knowledge: answers by big data are found through a process of automatic adjustment of data to models that do not carry any structural understanding beyond the solution of an *ad hoc* problem — blind computations bring no knowledge.

From the point of view of philosophy of mathematics, C. Calude and G. Longo[30] argue that there is a fundamental problem with the assumption that more data will necessarily yield more information: very large databases will *necessarily* contain arbitrary correlations due to the size, not the nature, of data. They show in the paper how combinatorial phenomena like Ramsey-type of correlations inexorably appear in all large enough databases. Those will be 'spurious correlation' (using the adjective 'spurious' in the sense of being false, although seeming to be genuine).

[26] **putnam1975a**.
[27] **wired:2008**.
[28] **wired:2008**.
[29] **napoletani:panza:struppa:2014**.
[30] **calude2017deluge**.

A well-known motto of Ramsey theory says that 'complete disorder is impossible', in the sense that upon increasing the cardinality of the samples some patterns will inevitably occur. This was confirmed when C. Di Prisco and myself introduced the *Principle of Ariadne*[31][32] , a new infinitary principle of set theory that is independent of the Axiom of Choice in **ZF** but which can be consistently added to the remaining axioms. This principle is related to Ramsey theory. Although the Principle of Ariadne is infinitary and big data works with huge (but finite) samples, it guarantees, similarly to the Axiom of Choice, the existence of certain structures in sets regardless to whether or not a rule for it can be specified.

Some reactions are already emerging

Some parts of the scientific establishment are starting to move. The European Union general data protection regulation 2016/679 (GDPR) (on effect in 2018) grants, in its Article 15, the "right of access by the data subject" [33]. The notion of interpretability in AI is the degree to which an observer can understand the cause of a decision.[34] In other words, people have the right to *understand* the reasoning and data that the model uses to arrive at a certain decision. The idea of justification — whether a model is accurate enough when tried against data — is not sufficient. Justification and interpretability are inversely related factors. The more accurate a model is, the less interpretable (and less understandable) it will be — and more like a black box. We have to consider that there is always a risk that extreme accuracy might not represent intelligence after all, but only a very effective way of dominion. As Pedro Domingos, in his *The Master Algorithm*,[35] puts it, "People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the world". In this way, without due diligence, ML might simply become an illegal activity. It cannot be avoided that the same type of diligence becomes standard for mathematics, and for science in general.

In his monograph *Superintelligence* Bostrom[36] addresses one of the most relevant questions regarding AI: would we want a superintelligence? As he says in Chapter 15, "The intelligence explosion might still be many decades off in the future. Moreover, the challenge we face is, in part, to hold on to our humanity: to maintain our groundedness, common sense, and good-humored decency even in the teeth of this most unnatural and inhuman problem". He compares the human perspective before a (perhaps inevitable) intelligence explosion to small children playing with a bomb. Superintelligence is a challenge for which we are not ready yet, and will not be ready for a long time.

As outlined in the section 'The quest for superintelligence', there are surprising connections between formal logic and superintelligence. Perhaps superintelligence will punish us for trying to frustrate it, or it may decide, in its almighty power, that a human race is not good enough to continue to exist. Can it cross the logic barriers? Who can guarantee that superintelligence will be content to use traditional logic?

From a completely different perspective, a group of mathematicians developed a computing problem that even the most intelligent machine learning algorithms will never be able to solve: it was proved[37] that not everything is knowable by machine learning. The algorithms are limited by

[31] **carnielli-diprisco:1993**.
[32] **carnielli-diprsico:2017**.
[33] `https://www.privacy-regulation.eu/en/.`
[34] **miller-explanat:2019**.
[35] **domingos2015master**.
[36] **bostrom:2014**.
[37] **ben2019learnability**.

the restrictions of mathematics, bounded by an unresolvability barrier similar to the incompleteness theorems Kurt Gödel developed in the 1930s. As a consequence, the containment problem is incomputable, but the practical relevance of this result should be contrasted with the arguments developed in[38] .

Although the limitative result by S.Ben-David et.al.[39] may remind an old objection pioneered by J.R. Lucas[40] and later reintroduced by R. Penrose,[41] it has a different meaning. Lucas appeals to the proof in Kurt Gödel's incompleteness theorem that Arithmetic cannot be proved consistent and complete, and interprets this as implying that no automated formal system could derive all of the truths of elementary arithmetic, at risk of being inconsistent. This led Lucas to conclude that the human mind will never be equated by a computation. Lucas's arguments, as well as Penrose's variations, are quite controversial and were rejected by many people[42][43][44] not because of what they conclude, but mainly for the insistence on blaming Gödel's incompleteness theorem by the inevitable gap that, according to them, there is between minds and machines[45].

The sort of mathematical revenge proposed by S. Ben-David et.al.[46] is from a different strain: it is an unsolvability barrier that promises that science, as we know it, will not be so easily replaced by correlations in huge databases — so long as we stay aware and alert enough. Perhaps this is the most humane way out of the dilemma.

Acknowledgements: I acknowledge support from the National Council for Scientific and Technological Development (CNPq), Brazil under research grants 307376/2018-4. I wish to thank Pedro Carrasqueira, Rafael Testa, Bruno Ramos Mendonça, Abilio Rodrigues and Pablo Rolim dos Santos for useful criticisms.

[38] **superintelligence2021**.
[39] **ben2019learnability**.
[40] **lucas1961minds**.
[41] **penrose:1989**.
[42] **laforte1998godel**.
[43] **bringsjord2000refutation**.
[44] **shapiro2003mechanism**.
[45] A survey of the complicated pros and cons against Lucas and Penrose's arguments can be found in `http://www.iep.utm.edu/lp-argue/`.
[46] **ben2019learnability**.