

Scientists *in silico*?

Carl McBride*

*Departamento de Ciencias y Técnicas Fisicoquímicas,
Facultad de Ciencias, Universidad Nacional de Educación a Distancia (UNED), 28040 Madrid, Spain.*

(Dated: November 2, 2017)

The end (for human scientists) is nigh? The posit of this discourse is that the majority, if not all, scientific research will eventually be undertaken by one, or a number of, weak artificial intelligences.

INTRODUCTION

“Hope we’re not just the biological boot loader for digital superintelligence. Unfortunately, that is increasingly probable” Elon Musk, twitter (August 3, 2014) [1]

The last three centuries have borne witness to spectacular progress in mathematics and the natural sciences, with developments such as calculus, classical mechanics, thermodynamics, quantum theory and general relativity, to name but a few. That said, in a classic paper by Eugene Wigner titled the ‘Unreasonable Effectiveness of Mathematics in the Natural Sciences’ [2] he observed that it is “. . . *not at all natural that laws of nature exist, much less that man is able to discover them*”. Towards the end of his paper he proposed the ‘empirical law of epistemology’ to account for our ability to understand, to such a surprising degree, the world around us. However, in the very long run it is perhaps inevitable that scientists will eventually become incapable of thinking openly enough to have ideas that have the potential to significantly advance science. In physics, for example, certain serious impasses seem to have been reached. Two of the most well known problems are; the long standing difficulty in melding together the somewhat disparate theories of general relativity and quantum mechanics; the last six decades have still not produced a satisfactory theory for quantum gravity, with competing theories like string theory or loop quantum gravity struggling to get the job done. Another problem that is causing headaches is the nature of so-called ‘dark matter’, non-Baryonic matter that is estimated to constitute a staggering 84% of the mass of the universe [3]. Theories to explain the observations of an almost solid-body like motion of galaxies originated with Fritz Zwicky applying the virial theorem in the 1930’s [4, 5]. Over eighty years later there is, as yet, no good theory to explain what is going on, with hypotheses like ‘weakly interacting massive particles’ (WIMPS), ‘modified Newtonian dynamics (MOND), and extensions to the Standard Model seemingly ruled out. There are now proposals for a ‘hidden sector’, composed of more complex self-interacting dark matter [6], an obscured section of the universe that we can only detect by way of gravity.

The Church-Turing-Deutsch principle states that: ‘every finitely realizable physical system can be perfectly simulated by a universal model computing machine operating by finite means’ [7]. Michael Nielsen has pointed out that *“No one has yet managed to deduce this form of Deutsch’s principle from the laws of physics. Part of the reason is that we don’t yet know what the laws of physics are!”* [8]. If this is indeed the case, and we understand less about physics than we think we do, then scientists may sooner-or-later end up instead dedicating their time to going down rabbit-holes; for example perhaps by unintentionally extrapolating beyond the validity of their models, or producing work that, although correct, is essentially irrelevant. Adopting the words of Nobel laureate David Gross, the issue *“. . . is not one of ideology but strategy: What is the most useful way of doing science?”* [9].

In Part I of this manuscript I shall focus on the practical aspects of the ongoing project of a complete knowledge of science: philosophy, our perceptive abilities, human error and biases, working practices, and aesthetics. In Part II I shall mention five examples of the progress that the (relatively) nascent field of artificial intelligence (AI) is making in science.

PART I

Philosophy of scientists

It was the illustrious Auguste Comte, founder of positivism and regarded as one of the first modern philosophers of science that, in 1835, famously stated: *“On the subject of stars, all investigations which are not ultimately reducible to simple visual observations are . . . necessarily denied to us. While we can conceive of the possibility of determining their shapes, their sizes, and their motions, we shall never be able by any means to study their chemical composition*

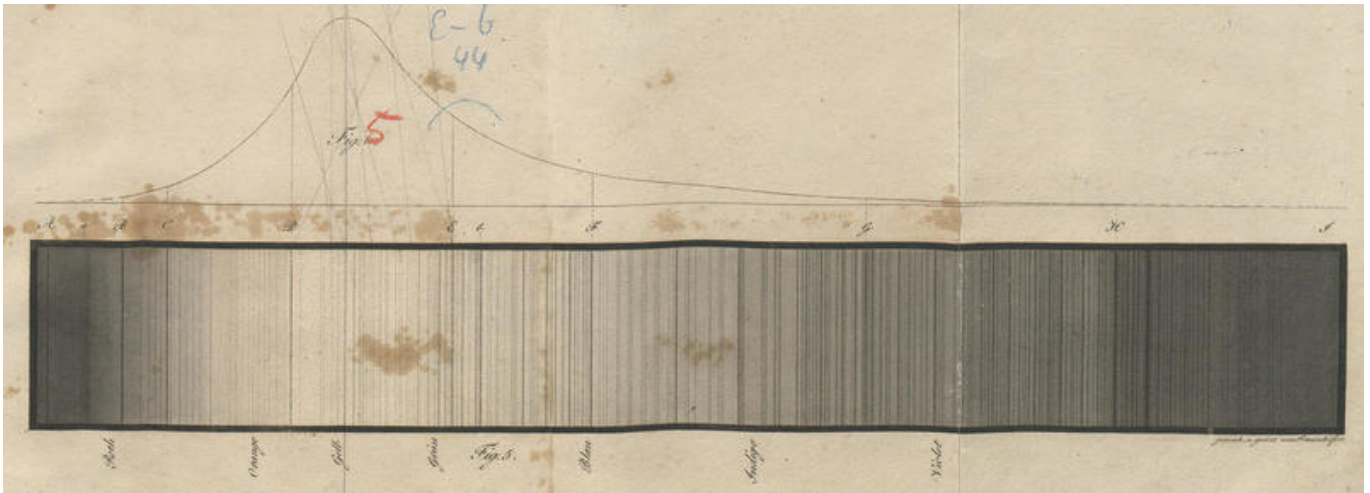


FIG. 1. Fraunhofer's adsorption lines in the solar spectrum, taken from his work published in 1817 [11].

or their mineralogical structure ... Our knowledge concerning their gaseous envelopes is necessarily limited to their existence, size ... and refractive power, we shall not at all be able to determine their chemical composition or even their density... I regard any notion concerning the true mean temperature of the various stars as forever denied to us." [10]. This statement was evidently made without being aware of the future impact of the experimental work of Joseph von Fraunhofer, published in 1817 [11] providing the basis for the eventual science of spectroscopy - the study of light spectra. In conjunction with the work of Gustav Kirchhoff on atomic absorption lines [12, 13], by the late 1800's it was indeed possible to have a very good idea of the chemical composition of stars. With the work of Josef Stefan it was also possible to obtain a value for the temperature of stars. The point is that one of the greatest intellects ever, by way of a reasoned argument, arrived at a conclusion that nature was able to throw back at us for being unfounded.

Incidents like this have unfortunately lead to there being some notable modern-day detractors of the value of philosophy in science; in a recent Google Zeitgeist conference, Stephen Hawking went as far as to say that 'philosophy is dead' [14]. The physicist Lawrence Krauss said "... the worst part of philosophy is the philosophy of science; the only people, as far as I can tell, that read work by philosophers of science are other philosophers of science. It has no impact on physics what so ever" [15]. Criticisms of the philosophy of science aside, we are interested in epistemology and the acquisition of knowledge. All scientists, be it consciously or unconsciously, have their own philosophy of science, which influences their approach to the work that they do. That said, it is often the case that working scientists do not have a clear premeditated 'philosophy' which guides them in their professional endeavours. It has been suggested that most scientists are 'critical realists', a subset of scientific realism [16]. Scientific realism is the viewpoint that there exists a world independently of ourselves (our minds), and that science does a progressively good job of describing both the 'observable' and the 'unobservable' parts of it. That science as a whole represents an ever improving approximate truth. An observable would be something one could directly experience, whereas an unobservable would be something for which there is indirect evidence, such as electrons, quarks etc. Within scientific realism there are three main variants: explanationist realism; entity realism; and structural realism. Explanationist realism believes in unobservables if they form part of a theory. entity realism asks for a stronger justification of unobservables before assigning them to be part of reality, such as being able to causally manipulate them, and structural realism, which believes in the structure, but does not assign reality to unobservables. As well as the realists, there are the anti-realists, with a dual viewpoint which can be divided up into empiricists, constructivists, operationalists or instrumentalists. In a survey in the magazine Physics World, undertaken by Robert P. Crease [17] it became evident that many scientists do not have a clear-cut philosophical approach to their work, and indeed a number of them simultaneously maintained traditionally diametric philosophies [18].

In Immanuel Kant's 'Critique of Pure Reason' (1787) he states: "All our knowledge begins with sense, proceeds thence to understanding, and ends with reason, beyond which nothing higher can be discovered in the human mind..."¹. The novelist F. Scott Fitzgerald famously wrote "the test of a first-rate intelligence is the ability to hold two opposed ideas

¹ Meiklejohn translation.



FIG. 2. A wild daffodil as seen with an ultraviolet filter (Photo courtesy of Bjørn Rørslett).

in the mind at the same time, and still retain the ability to function” [19]. An AI could hold an arbitrary number of opposed ideas in its memory at the same time, and test them one by one. If an AI can be programmed to reason better than ourselves, I see no reason why an AI could not make a better scientist. It is interesting to speculate whether an AI will be a naïve realist, having access to a Kantian ‘noumenal’ world, and the privilege of a vision of an objective reality, seeing *the truth, the whole truth, and nothing but the truth*.

Reality is an illusion: Interface theory of perception

“...humans could infer only as much as their senses allowed, but not experience the actual object itself”. Immanuel Kant [20]

Let us take for example the humble bee; bees, and many other birds and insects, can see ultraviolet light, and many flowers have a totally different aspect when seen in the ultraviolet (see Fig. 2). If bees saw as we do, they would almost certainly not survive very long, having great difficulty in locating their food source. Work by Donald Hoffman and co-workers on the Interface Theory of Perception [21] indicates that we are essentially incapable of correctly viewing reality. They find that *“veridical perceptions can be driven to extinction by non-veridical strategies that are tuned to utility rather than objective reality”* [22]. They postulate that evolution selects us on the basis of fitness for seeing the world from the point of view of survival, rather than rewarding ‘veridical perceptions’. In other words, our world-view is an internalised description, arrived at due to our animal nature. An analogy used is that of the graphical user interface on our computers; we see and interact with icons for files, CDs, printers, programs etc. We do not see the true nature of the files. If we could see the ‘true’ nature of the file, say the spiral of tiny distinctly orientated magnetic domains imprinted on a hard disk, we would be at a complete loss to get anything done. The reality of the string of 1’s and 0’s is hidden from us, and that is for the better; to be able to decipher the raw data would require a lot of time and energy, valuable resources when it comes to survival. Hoffman creates ‘fitness functions — mathematical functions that describe how well a given strategy achieves the goals of survival and reproduction’. These functions are then input into Monte Carlo simulations that run evolutionary ‘games’. The results seem to indicate that in the end we only see a ‘symbolic’ version of reality, because that is faster and cheaper and thus favors survival.

Human error: logarithms: exponentially difficult?

The difficulty associated with scientists doing science could be seen as being somewhat analogous to that of John Napier (as well as others to come). His tables of logarithms ‘Mirifici Logarithmorum Canonis Descriptio’ (1614) and ‘Logarithmorum Chilias Prima’ (1617) with Henry Briggs, were calculated by hand over a period of years. These tables were inevitably error-prone (for example see this analysis [23]), they had to be; the human mind is not conditioned to calculate error free to that extent. Such errors were not only limited to the calculations, but also introduced in the transcription and printing process.

A little over two-hundred years later Charles Babbage saw this weakness and invented his difference engine, a mechanical entity whose ‘world view’ is purely numerical. In 1822 he presented a short paper to the Astronomical Society of London ‘On the application of machinery to the computation of astronomical and mathematical tables’, and in 1831 he produced the book ‘Table of the Logarithms of the Natural Numbers from 1 to 108000’ [24]. Unfortunately

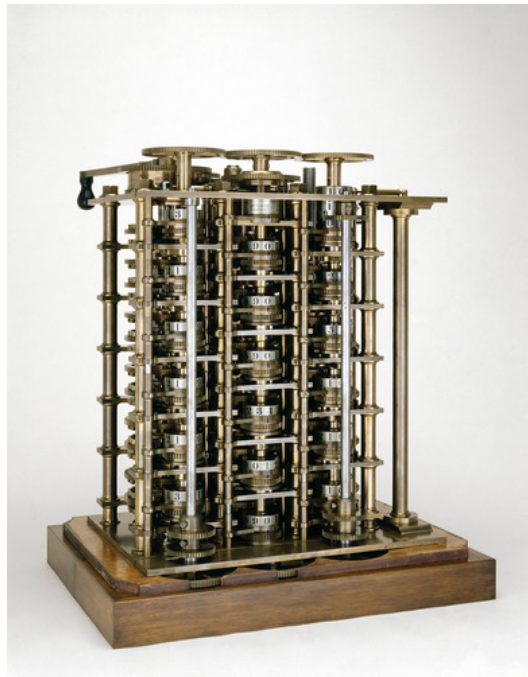


FIG. 3. A portion of Babbage's Difference Engine No. 1 (c. 1832) which currently stands in the Science Museum in London. (Source: Science Museum Group Collection Online).

only small sections of his Difference Engines were actually constructed, and his later analytical engine was never built, much to the chagrin of his sponsors, the British government.

Today, to attempt to calculate logarithms by hand would be considered a fool's errand. It is entirely possible that the current situation in the natural sciences is analogous; doing science 'by-hand' is inefficient and error-prone. What is needed this time is some sort of *intelligence engine*.

Experimenter bias

Psychologists have long been aware that we play host to a large array of cognitive biases [25], a number of issues have long been known, in particular a group of phenomena known as 'experimenter bias' –although it almost goes without saying that theoreticians can also partake. Due to this it can be sometimes exceedingly difficult to set-up a well designed experiment. In 1979 David Sackett compiled a 'catalogue of biases which may distort the design, execution, analysis and interpretation of research'. He identified no less than 35 factors that affect results [26]. A few examples of biases particularly relevant to scientific investigation are mentioned in an article by Regina Nuzzo [27]:

- Hypothesis myopia - where one finds what one is looking for, and ignores other hypothesis.
- The Texas sharpshooter - misinterpreting random patterns as being meaningful (the origin of the term is the story of a young boy who never misses; he simply shoots at the side of a barn, then *a posteriori* draws a target around the hole)
- Asymmetric attention - paying more attention to unusual results than to 'expected' results
- Just-so storytelling - *a posteriori* rationalisation of results
- Stopping data collection before time due to a false positive in A/B tests.

Various research protocols can put in place in order to avoid these problems by using mechanisms such as:

- Devils Advocacy - actively test alternative hypothesis
- Pre-commitment - publish intentions before performing data collection

- Teams of rivals - collaborate with groups that hold strongly different opinions
- Blind data analysis - undertake analysis on an ‘unknown’ data set, without knowing the results (see also [28])
- Having, and sticking to, stopping rules for A/B tests

Publish or perish

A less philosophical but more practical problem is summed up in the phrase ‘publish or perish’, which has been around since 1927 [29] and refers to the implicit pressure to publish academic studies in order to obtain tenure, funding, etc. Indeed, year on year there has been a significant rise in scientific production. The open access PubMed database currently indexes over 25 million journal articles in the fields of biomedical and the life sciences. Their annual statistical reports [30] indicate that in recent years this figure grows by almost a million per annum, corresponding to the addition of over one hundred papers an hour. It is inevitable that not all of this is necessarily quality work. It is also impossible, even for a whole team of researchers, to be constantly aware of all the research publication relevant to their work ². Correspondingly there is an ever increasing rate of retractions in the published scientific literature [31]. The situation is not just the fault of individual scientists. Recent work [32] has studied the enduring prevalence of false positives in scientific publications from the viewpoint of natural selection. To do this a dynamical population model was created whose protagonists had the ‘utmost integrity’, and never cheated ³. Their model consisted of three assumptions:

- Each laboratory of researchers were capable of positively identifying a true association
- Unless otherwise, the better this capability, the greater the rate of false positives
- If effort were to be expended in identifying false positives then productivity would decrease

One can then imagine how, in an environment where productivity means prizes, the eventual result is the indirect encouragement of ‘poor methodological practices’. In some fields, such as biomedical research, the situation has become so bad that it has been suggested that a good many of studies, for various reasons, are incorrect [33]. In a recent statistical analysis of publications in the field of anaesthesia [34] it was found that in a set of 5087 clinical trials published between 2000 and 2015, serious problems were identified in 4.1% of these papers, indicative of corrupted, incorrect, or just plain falsified data. These practical problems, (usually) unintentionally introduced by the scientists themselves, could be viewed as being, in some way, similar to the ‘transcription and printing’ problems encountered with the tables of logarithms mentioned earlier.

Mathematics and computer-assisted proofs: anti-aesthetic?

“A good mathematical proof is like a poem – this is a telephone directory!”

In this section we shall mention two mathematical problems that are easy to formulate, but after many years had passed, were eventually proven in good part with the aid of computers (Note: this is an anathema to many mathematicians). The above quip was made with regards to the proof of the four-color map problem, first posed in 1852 ‘four colours suffice to colour any planar map so that no two adjacent countries are the same colour’ [35, 36]. This was one of the first ever problems resolved with the help of a computer. The solution was eventually obtained in 1977, but only after using over 1,200 hours on an IBM 360 [37, 38], the resulting publication was over one-hundred pages long, and was accompanied by 465 pages of microfiche [39, 40]. Appel and Haken, in their popular article in Scientific American [41] interestingly observed that *“In a sense the program was demonstrating superiority not only in the mechanical parts of the task but in the intellectual areas as well”*.

Another high profile problem whose proof required the assistance of computers was the Kepler conjecture, which was first set down by Kepler himself in 1611 [42] in his treatise ‘Strena seu de nive sexangula’ (On the Six-Cornered Snowflake). The Kepler conjecture states that ‘no packing of congruent balls in Euclidean three-space has density

² Indeed there is now an AI powered search engine, [Semantic Scholar](#), designed to address this situation.

³ The protagonists in the study; simulated scientists, were all treated ethically.

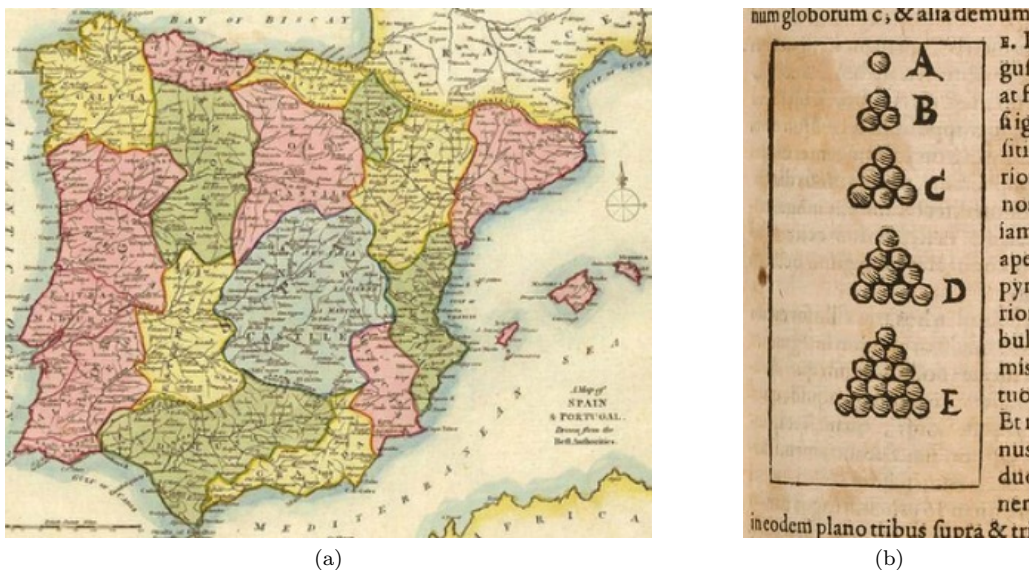


FIG. 4. (a) A map using four colours. (b) One of the diagrams in Kepler’s ‘Strena seu de nive sexangula’ (Courtesy of The Rare Book & Manuscript Library of the University of Illinois at Urbana-Champaign).

greater than that of the face-centered cubic packing’. The proof (by exhaustion) was finally published in 2006, in six sections, occupying a special issue of the journal, and spans over two hundred and sixty pages [43]. This proof was only formally verified in 2014 [44, 45] by what is known as the Flyspeck Project, which made use of the HOL (Higher Order Logic) Light proof assistant. However, if there ever was an example on a ‘non-surveyable proof’ it would be the solution of the Boolean Pythagorean triples problem; which takes up a 200 terabyte file [46]. A wonderful collection of other examples of computer-assisted mathematics can be found in Ref. [47].

In his 1940 essay ‘A Mathematician’s Apology’ G. H. Hardy extols:

The mathematician’s patterns, like the painter’s or the poet’s must be beautiful; the ideas like the colours or the words, must fit together in a harmonious way. Beauty is the first test: there is no permanent place in the world for ugly mathematics.

This requirement for a proof to be “...*elegant, concise and completely comprehensible by a human mathematical mind*” [41], or to be *like a poem*, a sentiment prevalent amongst mathematicians, has certainly paid off in the past, but is not a fundamental requirement, and under some circumstances, could conceivably hinder progress.

Our days are numbered

Our days are numbered, quite literally, in the form of probabilities. Studies have been undertaken to assess the chances that various sectors are susceptible to being ‘computerisable’ in the future. In a study in 2013 [48] ranking the probability of being computerisable, chemists and physicists, both involved in ‘creative intelligence tasks’ came in at ranks #173 and #175 respectively, with a 10% chance of being ‘computerisable’ at some point in time, eventually becoming redundant. Mathematicians fare a bit better, with a ranking of #135, and a 4.7% chance of being computerisable. All other physical scientists come in at #281, with a 43% chance.

Indeed, in a 2016 survey, taken amongst 352 of the attendees of two of the most important AI conferences, the consensus was reached that “*there is a 50% chance of AI outperforming humans in all tasks in 45 years and of automating all human jobs in 120 years*” [49].



FIG. 5. Arthur Samuel and his checkers playing IBM 701 (Photo courtesy of IBM).

PART II

Artificial Intelligence (AI)

“Shut up and calculate!” N. David Mermin [50]

The above quotation, although taken out of its original context, could well be the apothegm of narrow, or weak, AI. The philosopher John R. Searle, in his paper ‘Minds, brains, and programs’, subdivided artificial intelligence into ‘strong’ (general-purpose) and ‘weak’ [51]. He describes that in the case of a strong AI ... *“the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states.”* On the other hand, a weak AI is a program that has a specific, reduced remit. For example, a weak AI may be designed for image recognition; it could effectively and efficiently classify images into categories such as ‘car’, ‘flower’, ‘cat’ etc. A strong AI could do the same thing, but at the same time be thinking *‘I wonder how much that car costs?’*, *‘that is a nice flower!’*, *‘I am more of a dog than a cat person’* ...

In AI research there is currently no Master Algorithm [52], but rather there are five weak ‘camps’:

- Symbolists, using logic
- Evolutionaries with genetic programs [53]
- Bayesians, with graphical models [54]
- Analogisers, with support vectors
- Connectionists, with neural networks

as yet these threads have not synthesised into one universal approach, which could potentially be ‘strong’.

weak-AI and machine learning

Machine learning, sometimes known as statistical learning, is a sub-field of AI. The term ‘machine learning’ was coined by Arthur Samuel in 1959 in his paper describing a computer program designed to play the game of checkers [55]. In his seminal work he combined rote-learning with the groundbreaking ‘learning-by-generalisation’, leading to a program that could play a reasonable game after only 8-10 hours of running (on an IBM 701). He defined machine learning as when *“... computers [have] the ability to learn without being explicitly programmed”*. Computers playing games against humans has long been a testing ground for AI. For example, a significant milestone was achieved in

1997 when the IBM Deep Blue computer beat the world champion chess player Garry Kasparov [56]. In early 2011 IBM performed a live demonstration of a computer system, named Watson, designed to play the TV quiz show game *Jeopardy!*. The novel aspect of *Jeopardy!* is that rather than a standard Q&A format, usually with multiple-choice style answers, conversely the host presents somewhat cryptic answers, and the contestants must formulate the original question that leads to that answer. Watson played against two of the best previous winners of *Jeopardy!* and, in the third and final match, beat them both.

In 2016 a weak AI hit the headlines again, this time in the form of AlphaGo from Google DeepMind, a computer program [57] designed to play the ancient Chinese game of Go, a game that was previously thought to be, within the foreseeable future, essentially unplayable by a computer. To the amazement of many, AlphaGo beat 9th dan professional player Lee Sedol by four games to one in a competition consisting of five matches. More recently Ke Jie, considered the world's best player, was also beaten by AlphaGo, which won all three games [58]. The AlphaGo team have since dispensed with supervised learning, i.e. learning with some form of input from humans, and moved on to pure reinforcement learning. The latest iteration of the program, AlphaGo Zero [59] was capable, within three days of being turned on, to become good enough to beat the version that played Lee Sedol by 100 games to 0. Note that despite this amazing feat, AlphaGo Zero is a still weak AI; as it stands it would be completely incapable of playing chess without being reprogrammed by a human.

The Technological Singularity

It was the Polish mathematician Stanisław Ulam, in 1958 in his tribute to John von Neumann who mentions [60] *“the ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue.”*. In 1966 Irving John Good [61] described the following scenario: *“Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultraintelligent machine could design even better machines; there would then unquestionably be an “intelligence explosion,” and the intelligence of man would be left far behind. Thus the first ultraintelligent machine is the last invention that man need ever make...”*.

That said, perhaps the best argument against there ever being produced such an ultraintelligent machine in the first place was put forward by Bertram Bowden: *“there is no point in building a machine with the intelligence of a man, since it is easier to construct human brains by the usual method.”*.

Five examples of machine learning being used in science today

Deep learning: CERN

One of the ‘early adopters’ of the application of machine learning in scientific discovery was the high energy physics community [62], notably at the Conseil Européen pour la Recherche Nucléaire (CERN) where deep learning [63] is being used [64] to filter through the enormous data sets that are continuously being generated (up to 25 gigabytes per second! [65]) For example, looking for rare exotic particles amongst an enormous sea of particle collisions [66] such as the decay of the Higgs boson [67]. This can be viewed as a classification problem, well suited to deep learning. Indeed, *“...deep learning could even lead to the discovery of particles that no theorist has yet predicted”* [68].

Automated experiments: quantum mechanics and MELVIN

There is perhaps no branch of the physical sciences that is more universally recognised as being counterintuitive than quantum mechanics. It is a field that pushes our ability to reason, and thus our ability to understand, to the very limits of human intellectual capacity. In a recent paper [69] a computer program called MELVIN has been used to design configurations that, in the words of the authors *“...experiments found by our algorithm show a departure from conventional experiments in quantum mechanics in that they rely on highly unfamiliar, but perfectly conceivable experimental techniques”*. The algorithm starts with a ‘tool-box’ composed of commonly used optical components readily available in the laboratory such as prisms, mirrors, beam splitters etc. It then assembles these components randomly, and learns the output of this hypothetical laboratory setup. If the output is deemed to be ‘useful’, it is memorised and can come to form one of the building blocks of a subsequent, more elaborate setup. In other words,

the algorithm learns from experience. MELVIN, running for 150 hours (see [69] for details) designed 51 novel, yet feasible experiments. Without learning, the algorithm, running for a period of 250 hours, was unable to discover a number of the novelties found with learning.

Symbolic regression: Eureqa[®]

In 2009 Schmidt and Lipson created a computer program, now known as Eureqa[®] [70] that was able to re-discover important chunks of classical mechanics by itself. Directly quoting from the abstract in the Science paper: “*Without any prior knowledge about physics, kinematics, or geometry, the algorithm discovered Hamiltonians, Lagrangians, and other laws of geometric and momentum conservation. The discovery rate accelerated as laws found for simpler systems were used to bootstrap explanations for more complex systems, gradually uncovering the ‘alphabet’ used to describe those systems*” [71].

The Eureqa[®] code uses symbolic regression, a technique that is used not only to obtain the parameters for an equation from a data set (i.e. traditional regression analysis), but also the equation(s) themselves by way of an evolutionary algorithm [53]. The evolutionary algorithm creates trial models by mixing and matching mathematical ‘genes’ (operators, functions etc) to form a ‘creatures’ (equations) well suited to its environment (the data set).

Artificial neural networks: molecular potentials

In the computer simulation of liquids one has the situation where, lets take as an example the molecule water, there are at least one hundred and thirty thermodynamic models currently being studied in the literature [72] a good number of which are built upon, or extensions of, the parameterised Lennard-Jones model in conjunction with point charges. In the publication [73] instead neural network potentials, a technique originally designed for brain research, were used as a set of ‘very flexible functions’, that were trained with the results of a range of condensed phase configurations in order to ‘learn’ the *ab-initio* potential energy surface of water molecules. Such a model is not pre-biased by any prior conceptions of what, in this case a water molecule, should ‘look-like’, no matter how valid the physical reasoning behind the model is. This is not the only example, artificial neural network potentials have also been successfully developed for Al³⁺ ions dissolved in water [74], aqueous NaOH solutions [75], silicon [76], gold nanoparticles [77], as well as a number of other systems [78, 79]. Jörg Behler has written some very good tutorials as to how to implement molecular potentials in the form of neural networks [80, 81].

Monte Carlo tree search: organic molecule synthesis

AI is being applied to the field of organic chemistry. Retrosynthetic analysis involves proposing a synthesis route by working backwards to molecules that one knows how to make. The researchers trained their AI [82] with the Reaxys[®] database, which contains over 40 million chemical reactions, obtained from patents and publications spanning dating back to 1771. Using a combination of Monte Carlo tree search in conjunction with a deep neural network, they tested their program on 40 randomly selected molecules, finding a retrosynthesis route for 95% of time, beating the state of the art Best-First Search, using hand-coded heuristics, which provided routes for only 22.5% of the molecules in the allotted time.

Possible limitations of AI

It is most likely the case that to do science requires a strong AI, something that some people suggest may be impossible to create in a computer. In Searle’s aforementioned paper [51] he puts forth an argument as to why strong AI is not possible. Searle created the Chinese room scenario, which builds upon the famous Turing test [84]. Turing’s game essentially involved convincing a human interrogator that he or she is asking questions of a real person and not a computer program, adding that “...*in order that tones of voice may not help the interrogator the answers should be written, or better still, typewritten. The ideal arrangement is to have a teleprinter communicating between the two rooms.*” In the Chinese room *gedankenexperiment*, Searle supposes that:

- There is a computer locked in a room that, when fed information written in Chinese, responds satisfactorily via a printout, again in Chinese, in the process easily passing the Turing test with its answers.

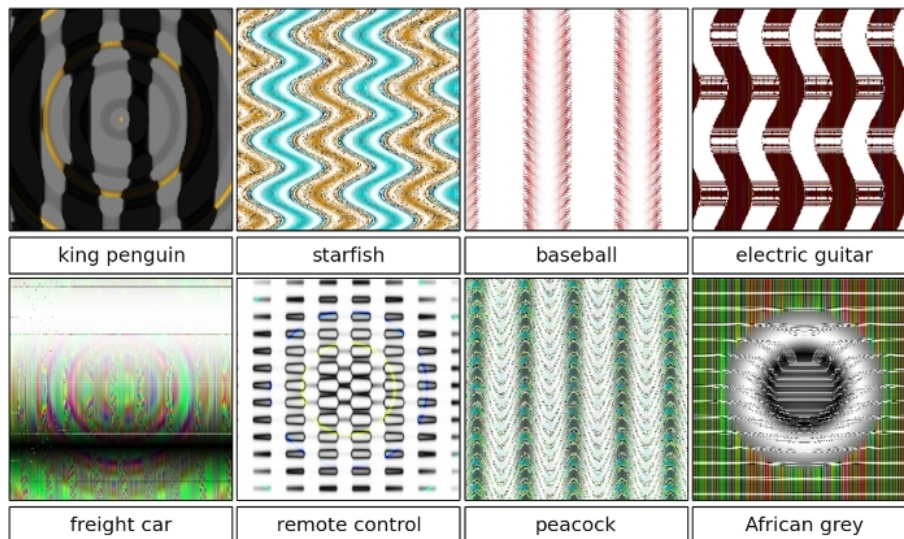


FIG. 6. Examples of an AI being fooled into annotating unrecognisable images with extremely high confidence (Source: Ref. [83]).

- The computer, being a computer, follows a deterministic algorithm
- Given time and patience, Dr. Searl could sit in the room instead of the computer, and also follow the algorithm (provided for him in English), and also output satisfactory answers in Chinese
- Dr. Searl has absolutely no knowledge of Chinese whatsoever, and thus no understanding of either what the input nor the output means
- Therefore, the computer also has no real understanding of what it is doing, and as such does not think and has no mind of its own.

There have been numerous publications both for and against this argument⁴. With regards to ‘real understanding’ Richard Feynman once said about one of the best theories that physicists have developed to date, “... *I think I can safely say that nobody understands quantum mechanics.*” [85]. We may not understand quantum mechanics, but nobody can deny that we certainly spend an inordinate amount of time *thinking* about it! All said and done, as we have seen with Comte, philosophical arguments can sometimes be overtaken by practical advances. If the Church-Turing-Deutsch principle [7] does hold, then there should be no reason why a strong AI cannot be built. Given that a strong AI in principle can be (and therefore eventually will be) built, here we mention some of the current problems in AI.

- The statistical approach of machine learning may not be enough in itself to derive the laws of physics and apply them to novel situations. Symbolic learning, using logic, will probably need to form part of the AI’s make-up in order to have the ability of relational reasoning [86, 87].
- AI’s are not infallible. For example, a task that deep neural networks have become particularly good at is image recognition. Each year ImageNet Large-Scale Visual Recognition Challenge is held to test AI’s ability to perform various tests on a large database of images, and each year the results improve. In 2014, in the image classification part of the challenge the best AI achieved a 6.66% error rate [88], rivalling human annotators. However, it has been shown that deep neural nets can produce false positives from images (see Fig. 6) that are meaningless to humans [83].
- It would be difficult to define a generalised ‘reward function’. One cannot simply say to an AI ‘do science’, it needs to know when it is making progress. It somehow needs to embody, or more precisely encode, the question that lies at the heart of science: ‘Why?’

⁴ Google Scholar lists over 5,800 citations of the original paper

- Neural networks, especially deep neural networks, have been accused of being black-boxes, providing wonderful results, but they themselves are seemingly indecipherable, the so-called ‘interpretability problem’. However, that may be changing, with theories such as the ‘information bottleneck’ method [89, 90], along with a multitude of mechanisms for teasing out the features that the deep neural network has honed in on as being the salient features of, say, an image [91, 92].

EPILOGUE

Artificial intelligence is ever increasingly finding its way into the scientists tool-box as a powerful technique to aid discovery. Although there is much ongoing work to develop AIs that perform certain human tasks better than humans can, the biggest rewards will almost certainly come from AIs performing unthought-of tasks in an unforeseen manner. However, asking an AI to perform science *per se* is a long way off, with major advances required before any AI can encapsulate aspects such as motivation and independent creativity, required to tackle such a monumental task. That said, eventually we may find ourselves leaving all scientific research to the AI’s. The predicted ‘Fourth Industrial Revolution’ [93] could also bring with it a revolution in the way we undertake science. When will this ‘digital superintelligence’ that can conduct unsupervised science be developed? According to the Maes-Garreau law⁵, sometime in the next 20 years. Until then we will just have to learn to cooperate with each other [94].

*... “Forty-two!” yelled Loonquawl. “Is that all you’ve got to show for seven and a half million years’ work?”
– “I checked it very thoroughly,” said the computer, “and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you’ve never actually known what the question is.”*

Douglas Adams - The Hitchhiker’s Guide to the Galaxy

Acknowledgments

The author would like to thank Cristina Santa Marta Pastrana and Juan J. Freire for their support during the writing of this manuscript. This work has been supported by a Universidad Nacional de Educación a Distancia (UNED) Postdoctoral Grant (2013-018-UNED-POST).

* carl.mcbride@ccia.uned.es; Dr.C.McBride@gmail.com

- [1] Elon Musk, *twitter* (3 August 2014).
- [2] Eugene P. Wigner, “Unreasonable effectiveness of mathematics in the natural sciences,” *Communications on Pure and Applied Mathematics* **13**, 1–14 (1960).
- [3] C. L. Bennett et al, “Nine-year wilkinson microwave anisotropy probe (WMAP) observations: Final maps and results,” *The Astrophysical Journal Supplement Series* **208**, 20 (2013).
- [4] Fritz Zwicky, “Die rotverschiebung von extragalaktischen nebeln,” *Helvetica Physica Acta* **6**, 110 (1933).
- [5] Fritz Zwicky, “On the masses of nebulae and of clusters of nebulae,” *Astrophysical Journal* **86**, 217 (1937).
- [6] David N. Spergel and Paul J. Steinhardt, “Observational evidence for self-interacting cold dark matter,” *Physical Review Letters* **84**, 3760 (2000).
- [7] David Deutsch, “Quantum theory, the Church-Turing principle and the universal quantum computer,” *Proceedings of the Royal Society A: Mathematical, Physical & Engineering Sciences* **400**, 97–117 (1985).
- [8] Michael Nielsen, “The physical origin of universal computing,” *Quanta Magazine* **October 27** (2015).
- [9] Natalie Wolchover, “A fight for the soul of science,” *Quanta Magazine* **December 16** (2015).
- [10] Auguste Comte, *Cours de philosophie positive: La philosophie astronomique et la philosophie de la physique* (Bachelier, 1835).
- [11] Joseph von Fraunhofer, “Bestimmung des brechungs- und farbenzerstreuungs-vermögens verschiedener glasarten, in bezug auf die vervollkommnung achromatischer fernröhre,” *Denkschriften der Königlichen Akademie der Wissenschaften zu München* **5**, 193 (1817).
- [12] Gustav R. Kirchhoff, “Ueber das verhältniss zwischen dem emissionsvermögen und dem absorptionsvermögen der körper für wärme und licht,” *Annalen der Physik* **185**, 275–301 (1860).

⁵ Maes-Garreau law: the amusing observation that these type of predictions always seem to coincide with the number of years until the retirement age of the person making it.

- [13] Gustav R. Kirchhoff, "I. On the relation between the radiating and absorbing powers of different bodies for light and heat," *Philosophical Magazine Series 4* **20**, 1–21 (1860).
- [14] <https://www.zeitgeistminds.com/talk/60/unified-theory-professor-stephen-hawking>.
- [15] Ross Andersen, "Has physics made philosophy and religion obsolete?" *The Atlantic* **April 23** (2012).
- [16] John Polkinghorne, *Faith, Science and Understanding* (Yale University Press, 2001).
- [17] Robert P. Crease, "Whats your philosophy?" *Physics World* **14**, 18 (October 2001).
- [18] Robert P. Crease, "This is your philosophy," *Physics World* **15**, 15 (April 2002).
- [19] F. Scott Fitzgerald, *The Crack-Up* (Esquire, February 1936).
- [20] Immanuel Kant, *De Mundi Sensibilis atque Intelligibilis Forma et Principiis*, Inaugural dissertation, University of Königsberg (1770).
- [21] Donald D. Hoffman, Manish Singh, and Chetan Prakash, "The interface theory of perception," *Psychonomic Bulletin & Review* **22**, 1480–1506 (2015).
- [22] Justin T. Mark, Brian B. Marion, and Donald D. Hoffman, "Natural selection and veridical perceptions," *Journal of Theoretical Biology* **266**, 504–515 (2010).
- [23] <http://www.pmonta.com/tables/logarithmorum-chilias-prima/index.html>.
- [24] Charles Babbage, *Table of the logarithms of the natural numbers, from 1 to 108000* (London, B. Fellowes, 1831).
- [25] David McRaney, *You Are Not So Smart* (Gotham Books, 2011).
- [26] David L. Sackett, "Bias in analytic research," *Journal of Chronic Diseases* **32**, 51–63 (1979).
- [27] Regina Nuzzo, "How scientists fool themselves - and how they can stop," *Nature* **526**, 182–185 (2015).
- [28] Robert MacCoun and Saul Perlmutter, "Blind analysis: Hide results to seek the truth," *Nature* **526**, 187–189 (2015).
- [29] Clarence Marsh Case, "Scholarship in sociology," *Sociology and Social Research* **12**, 325 (1927).
- [30] <https://www.nlm.nih.gov/bsd/licensee/baselinestats.html>.
- [31] Richard Van Noorden, "Science publishing: The trouble with retractions," *Nature* **478**, 26–28 (2011).
- [32] Paul E. Smaldino and Richard McElreath, "The natural selection of bad science," *Royal Society Open Science* **3**, 160384 (2016).
- [33] John P. A. Ioannidis, "Why most published research findings are false," *PLOS Medicine* **2**, e124 (2005).
- [34] John B. Carlisle, "Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals," *Anaesthesia* **72**, 944–952 (2017).
- [35] John Mitchem, "On the history and solution of the four-color map problem," *The Two-Year College Mathematics Journal* **12**, 108–116 (1981).
- [36] Robin Wilson, *Four Colors Suffice: How the Map Problem Was Solved* (Princeton University Press, 2004).
- [37] Kenneth Appel and Wolfgang Haken, "Every planar map is four colorable. Part I: Discharging," *Illinois Journal of Mathematics* **21**, 429–490 (1977).
- [38] Kenneth Appel, Wolfgang Haken, and J. Koch, "Every planar map is four colorable. Part II: Reducibility," *Illinois Journal of Mathematics* **21**, 491–567 (1977).
- [39] Kenneth Appel and Wolfgang Haken, "Microfiche supplement to "Every planar map is four colorable. Part I and Part II",," *Illinois Journal of Mathematics* **21** (1977).
- [40] Kenneth Appel and Wolfgang Haken, "Microfiche supplement to "every planar map is four colorable",," *Illinois Journal of Mathematics* **21** (1977).
- [41] Kenneth Appel and Wolfgang Haken, "The solution of the four-color-map problem," *Scientific American* **237**, 108–121 (1977).
- [42] Johannes Kepler, *Strena seu de nive sexangula* (Frankfurt: Gottfried. Tampach, 1611).
- [43] Thomas C. Hales and Samuel P. Ferguson, "The Kepler conjecture," *Discrete & Computational Geometry* **36**, 5–265 (2006).
- [44] Thomas Hales, Mark Adams, Gertrud Bauer, Dat Tat Dang, John Harrison, Truong Le Hoang, Cezary Kaliszyk, Victor Magron, Sean McLaughlin, Thang Tat Nguyen, Truong Quang Nguyen, Tobias Nipkow, Steven Obua, Joseph Pleso, Jason Rute, Alexey Solovyev, An Hoai Thi Ta, Trung Nam Tran, Diep Thi Trieu, Josef Urban, Ky Khac Vu, and Roland Zunkeller, "A formal proof of the Kepler conjecture," *arXiv*, 1501.02155 (2015).
- [45] <https://github.com/flyspeck/flyspeck>.
- [46] Evelyn Lamb, "Two-hundred-terabyte maths proof is largest ever," *Nature* **534**, 17–18 (2016).
- [47] David H. Bailey and Jonathan M. Borwein, "Computer-assisted discovery and proof," *Contemporary Mathematics* **457**, 21–52 (2008).
- [48] Carl Benedikt Frey and Michael A. Osborne, "The future of employment: How susceptible are jobs to computerisation?" *Oxford Martin School Publications* (17 September 2013).
- [49] Katja Grace, John Salvatier and Allan Dafoe, Baobao Zhang, and Owain Evans, "When will AI exceed human performance? Evidence from AI experts," *arXiv*, 1705.08807 (2017).
- [50] N. David Mermin, "What's wrong with this pillow?" *Physics Today* **42**, 9 (April 1989).
- [51] John R. Searle, "Minds, brains, and programs," *Behavioral and Brain Sciences* **3**, 417–424 (1980).
- [52] Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (Basic Books, 2015).
- [53] Stephanie Forrest, "Genetic algorithms: principles of natural selection applied to computation," *Science* **261**, 872–878 (1993).
- [54] Zoubin Ghahramani, "Review: Probabilistic machine learning and artificial intelligence," *Nature* **521**, 452–459 (2015).
- [55] Arthur Lee Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*

- opment **3**, 210–229 (1959).
- [56] Garry Kasparov, *Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins* (PublicAffairs, 2017).
- [57] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis, “Mastering the game of Go with deep neural networks and tree search,” *Nature* **529**, 484–489 (2016).
- [58] <https://events.google.com/alphago2017/>.
- [59] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis, “Mastering the game of go without human knowledge,” *Nature* **550**, 354–359 (2017).
- [60] Stanisław Ulam, “John von Neumann 1903-1957,” *Bulletin of the American Mathematical Society* **64**, 1 (1958).
- [61] Irving John Good, “Speculations concerning the first ultraintelligent machine,” *Advances in Computers* **6**, 31–88 (1966).
- [62] B. Denby, “Neural networks and cellular automata in experimental high energy physics,” *Computer Physics Communications* **49**, 429–448 (1988).
- [63] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, “Review: Deep learning,” *Nature* **521**, 436–444 (2015).
- [64] Nicola Jones, “Computer science: The learning machines,” *Nature* **505**, 146–148 (2014).
- [65] <http://cds.cern.ch/record/1997399>.
- [66] P. Baldi, P. Sadowski, and D. Whiteson, “Searching for exotic particles in high-energy physics with deep learning,” *Nature Communications* **5**, 4308 (2014).
- [67] P. Baldi, P. Sadowski, and D. Whiteson, “Enhanced higgs boson to $\tau^+\tau^-$ search with deep learning,” *Physical Review Letters* **114**, 111801 (2015).
- [68] Davide Castelvecchi, “Artificial intelligence called in to tackle LHC data deluge,” *Nature* **528**, 18–19 (2015).
- [69] Mario Krenn, Mehul Malik, Robert Fickler, Radek Lapkiewicz, and Anton Zeilinger, “Automated search for new quantum experiments,” *Physical Review Letters* **116**, 090405 (2016).
- [70] <http://www.nutonian.com/products/eureka/>.
- [71] Michael Schmidt and Hod Lipson, “Distilling free-form natural laws from experimental data,” *Science* **324**, 81–85 (2009).
- [72] http://www.sklogwiki.org/SklogWiki/index.php/Water_models.
- [73] Tobias Morawietz, Andreas Singraber, Christoph Dellago, and Jörg Behler, “How van der Waals interactions determine the unique properties of water,” *PNAS* **113**, 8368–8373 (2016).
- [74] Helmut Gassner, Michael Probst, Albert Lauenstein, and Kersti Hermansson, “Representation of intermolecular potential functions by neural networks,” *Journal of Physical Chemistry A* **102**, 4596 (1998).
- [75] Matti Hellström and Jörg Behler, “Structure of aqueous NaOH solutions: insights from neural-network-based molecular dynamics simulations,” *Physical Chemistry Chemical Physics* **19**, 82 (2017).
- [76] Ekin D. Cubuk, Brad D. Malone, Berk Onat, Amos Waterland, and Efthimios Kaxiras, “Representations in neural network based empirical potentials,” *Journal of Chemical Physics* **147**, 024104 (2017).
- [77] Siva Chiriki, Shweta Jindal, and Satya S. Bulusu, “Neural network potentials for dynamics and thermodynamics of gold nanoparticles,” *Journal of Chemical Physics* **146**, 084314 (2017).
- [78] Sönke Lorenz, Axel Groß, and Matthias Scheffler, “Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks,” *Chemical Physics Letters* **395**, 210–215 (2004).
- [79] Sergei Manzhos, Xiaogang Wang, Richard Dawes, and Tucker Carrington Jr., “A nested molecule-independent neural network approach for high-quality potential fits,” *Journal of Physical Chemistry A* **110**, 5295–5304 (2006).
- [80] Jörg Behler, “Constructing high-dimensional neural network potentials: A tutorial review,” *International Journal of Quantum Chemistry* **115**, 1032–1050 (2015).
- [81] Jörg Behler, “Perspective: Machine learning potentials for atomistic simulations,” *Journal of Chemical Physics* **145**, 170901 (2016).
- [82] Marwin Segler, Mike Preuß, and Mark P. Waller, “Towards “AlphaChem”: Chemical synthesis planning with tree search and deep neural network policies,” *arXiv* , 1702.00020 (2017).
- [83] Anh Nguyen, Jason Yosinski, and Jeff Clune, “Deep neural networks are easily fooled: High confidence predictions for unrecognizable images,” *arXiv* , 1412.1897 (2015).
- [84] Alan M. Turing, “Computing machinery and intelligence,” *Mind* **59**, 433–460 (1950).
- [85] Richard P. Feynman, *The Character of Physical Law* (MIT Press, 1967).
- [86] Adam Santoro, David Raposo, David G.T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap, “A simple neural network module for relational reasoning,” *arXiv* , 1706.01427 (2017).
- [87] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià Puigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis, “Hybrid computing using a neural network with dynamic external memory,” *Nature* **538**, 471–476 (2016).
- [88] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, “ImageNet large scale visual recognition challenge,” *International Journal of Computer Vision* **115**, 211–252 (2015).
- [89] Naftali Tishby, Fernando C. Pereira, and William Bialek, “The information bottleneck method,” *arXiv physics/0004057* (2000).
- [90] Ravid Shwartz-Ziv and Naftali Tishby, “Opening the black box of deep neural networks via information,” *arXiv* , 1703.00810 (2017).

- [91] Davide Castelvetti, “Can we open the black box of AI?” *Nature* **538**, 20–23 (2016).
- [92] Paul Voosen, “How AI detectives are cracking open the black box of deep learning,” *Science* (2017), 10.1126/science.aan7059.
- [93] Klaus Schwab, *The Fourth Industrial Revolution* (Crown Business, 2017).
- [94] Jacob W. Crandall, Mayada Oudah, Tennom, Fatimah Ishowo-Oloko, Sherief Abdallah, Jean-François Bonnefon, Manuel Cebrian, Azim Shariff, Michael A. Goodrich, and Iyad Rahwan, “Cooperating with machines,” *arXiv* , 1703.06207 (2017).