A Criticism of the Technological Singularity

Alexander K. Seewald[1]

Seewald Solutions, Lärchenstraße 1, 4616 Weißkirchen a.d. Traun, Austria,
alex@seewald.at,
WWW home page: www.seewald.at

Abstract. The compelling story of self-improving machines which become exponentially smarter up to inconceivable intelligence has inspired some of the best science fiction literature [1–3], but is also taken seriously by many researchers. This story is however based on empirical observations of seemingly exponential processes such as Moore's law in the semiconductor industry, and contains multiple fallacies concerning self-improvement of intelligent systems (including humans), which upon close look are implausible. Deep Learning has been heralded as a major step in this direction, however a closer look again shows many open issues with this approach, leading us to conclude that we deciphered only a small part of one method which nervous systems *may* use to create intelligent behaviour; that seemingly simple tasks like image classification and segmentation are still AI-complete; and that true Artificial General Intelligence (AGI) still lies at least several centuries in the future. But even an AGI would not be able to exponentially self-improve without further advances. These fallacies abound in science fiction literature as well as in scientific papers, and we will illustrate our analysis with appropriate examples.

Keywords: Singularity, Criticism, Artificial General Intelligence, Moore's Law, Cybernetics, AI History

1 INTRODUCTION

The technological singularity, or Intelligence Explosion Hypothesis (IEH), is a theoretical concept where exponential improvements in intelligence (e.g. a factor of 2 every year) yield corresponding speedups of the underlying system where these speedups were obtained, making the next exponential improvement step by the same system exponentially faster (e.g. a factor of 2 every 6 months, every 3 months, every 1.5 months, ...), thus yielding exponential intelligence improvement steps in exponentially shorter time, rapidly reaching a point where the total intelligence of such a system reaches a mathematical singularity (e.g. $\frac{1}{0}$). Self-recursive improvement is thus an essential feature of this concept. The effect is in each case accelerated technological change, leading to a rapid transition to a state where the human condition would be challenged, and the future would be unpredictable.

The concept is not new. Already in the 1950ies, John von Neumann [4] claimed that accelerating technological progress was approaching a mathematical singularity beyond which human affairs would radically change.[1] Good [5] postulated an intelligence explosion by recursive self-improvement of a machine intelligence, which he called an *ultraintelligent* machine. Lem [6] published a science fiction novel about a military AI that recursively increases its intelligence repeatedly, rapidly approaching just such a singularity. Two years later, Vinge [7] contended that a rapidly self-improving AI would quickly approach a technological singularity beyond which reality would be unpredictable, and later elaborated on this in [8], where he predicted the singularity to happen between 2005 and 2030. Solomonff [9] described six AI milestones he believed would lead to AI approaching practically infinite intelligence in finite time, assuming again a self-recursive improvement (in the form of: having computers makes it cheaper to create newer, more powerful computers) and a constant expenditure on exponentially decreasing hardware costs. Moravec [10] predicted that in 2038 – due to Moore's law and similar trends – robotic reasoning and behavior would exceed that of humans. In 1990, examining evidence outside robotics, Kurzweil [11, 12] came to similar conclusions and put the singularity at around 2045. Yudkowsky [13] differentiates three different singularity schools:

1. *Accelerating Change:* Technological change follows (often exponential or super-exponential) smooth curves, so the arrival of new technologies – or when they cross key thresholds (such as Artificial General Intelligence) – can be predicted with reasonable precision.
2. *Event Horizon:* To predict a superintelligence's plans, you need to be at least as smart yourself. Thus the future after the creation of superintelligence is completely unpredictable. This is most similar to the definition by Vinge in [8].
3. *Intelligence Explosion:* The positive self-improvement feedback cycle of intelligence does not stop but triggers further and ever faster intelligence improvements of similar magnitude and creates a superintelligence before it hits physical limits.

He notes that their claims contradict one another: for example, if we use Moore's law to predict computing performance in 2099, we contradict both *Event Horizon* (which state that we cannot know the future after superintelligences) and *Intelligence Explosion* (because progress will run faster once smarter-than-human minds and nanotechnology are integrated into the process). However these comprehensive discussions do not at all indicate that any of these schools describe even remotely plausible scenarios.

Sandberg [14] compares models of the technological singularity and concludes that mathematical models for growth exhibit at least exponential growth as this is the signature of linear self-coupling terms. He adds that if efficiences of scale exist, superexponential growth or finite time singularities appear to be generic. He notes that mathematical singularities are likely indicators for transitions to other domains of growth, or that unmodeled factors will become relevant close to the point and are not meant to be taken at face value. He also notes that the most solid finding is that even small increasing returns in a growth model can produce radical growth, and that if mental capital (embodied in humans, artificial intelligence or posthumans) becomes relatively cheaply copyable, extremely rapid growth is likely to follow. Concluding, he notes that there is

[1] This was at a time of excessive optimism in AI research, similar to the present (2021).

a lack of models of how an intelligence explosion could occur and that available evidence show that human experts are usually weak at long-term forecasting even without apparent singularities.

There are two distinct ways such a singularity is presumed to happen:

1. *Hard takeoff*: Runaway intelligence, recursive speeding-up self-improvement *in the blink of an eye*.
2. *Soft takeoff*: An exponential speedup similar what we have experienced due to Moore's law.

As we will see, a hard takeoff is extremely unlikely given the very strict requirements; however even a soft takeoff over a reasonable timeframe of 10-20 years also seems rather unlikely. A sufficiently slow soft takeoff taking several centuries would be undistinguishable from the status quo and we will therefore not address it.

There are many issues with past work on this topic, so we will first only mention a few of them: For example, above work mostly assumes that increases in research funding – one working example of a self-improving system – yield exponential increases in research progress while these are actually sublinear [15]. This is also supported by the finding of Hanson [16] who found that past progress in research is not perceived as increasing over time by the researchers themselves. In fact, based on the much more reliable past rate of progress estimated by researchers themselves, Hanson [17] (p.61–63) estimates at least two to four centuries until an Artificial General Intelligence – i.e. a human-level AI – becomes available.

It is sometimes claimed (e.g. in [8]) that intelligence amplification in humans could yield comparable results. However it seems obvious that the speedups necessary for a runaway intelligence explosion are not applicable to biological nervous system since the necessary self-improvement loop is not feasible for biological entities, at least at ever decreasing temporal lengths. Therefore we do not address this topic here.

Lastly, it should be noted that from an evolutionary standpoint, our facility for rational human thought has not evolved for drawing valid and statistically accurate conclusions, but rather to impress and persuade other humans by telling compelling stories [18]. So it does not come as a surprise that the singularity – a story of self-improving machines with unlimited potential – has a wide range of proponents as well as detractors. Please bear this in mind for the following discussion.

2 SINGULARITY REQUIREMENTS

There are three key requirements for any kind of technological singularity.

1. Recursive self-improvement must be feasible, regardless of the intelligence level to start with. It is not at all clear that recursive self-improvement is actually possible. If we go from the most intelligent system known so far – humans – cognitive introspection is a bad indicator of how things are actually implemented in our brains, and can only barely inform improvements in intelligence. Systematic experiments were and are still needed to determine good ways to improve mathematical abilities, memory, and logic even temporarily [19]. General intelligence is practically

not improvable.[2] In machine learning after decades of research, there seem to be no true examples of unsupervised self-learning systems that automatically improve without explicit or implicit external feedback. In each case where machines have seemingly improved themselves (e.g. speech recognition, faster CPUs and GPUs, image classification and segmentation, self-driving cars, spam filtering, ...), this process was enabled by human ingenuity at exponentially increasing scales of effort and cost.

2. Recursive self-improvement must be feasible at a cost at most linear in the achieved improvement. For example, to make a system twice as smart, we must incur at most a linear cost $O(x)$ with x corresponding to the initial smartness of the system. So making a system x times smarter must be only x times as hard.[3] Otherwise the maximum achievable intelligence grows much more slowly, or even converges asymptotically to a limit never to be exceeded – surely the opposite of a singularity. However indications abound that it is actually much harder to make a system smarter, on the order of $O(x^2)$ or even $O(e^x)$. For example, increasing research funding yields much less than linear increases in research progress [15], indicating that costs are much more than linear w.r.t progress. In fact vulnerabilities such as Meltdown and Spectre [22] clearly indicate that previous costs at least in computer chip design have been underestimated in that they did not account for very subtle significant errors in previous chip designs.
3. At a linear cost of recursive self-improvement, we would still have to have an additional exponential process to achieve the necessary speedup for a singularity with *hard take-off*. For the *soft take-off* scenario the previous two requirements are sufficient. For example, we could exponentially increase the available budget for such a project, or a system twice as intelligent could only need half the cost to improve by another factor of two.[4] Another usual candidate is the exponentially increasing number of transistors in microchips at shrinking costs due to Moore's law. However this relies on 1) that is it possible to parallelize the recursive self-improvement process to an arbitrary number of subprocesses which can all run in parallel, 2) for the exponential process to hold at least until the earliest timepoint where the singularity can happen. Both assumptions are doubtful: 1) ignores all communication costs between processors, but these also increase superlinear with the number of processors[5]; and 2) relies heavily on Moore's law which will hold only for another one to two decades at most (see Section 3).

The most critical requirement is the second one: Achieving the singularity assumes a sublinear effort for intelligence improvement. E.g. to achieve double intelligence, a constant effort is needed regardless of the previous intelligence level. Even a slight exponent on the effort will prevent a singularity from happening. Naam [23] has argued against such a superexponential runoff. He has pointed out that we already see recursive

[2] Stumpf [20] finds correlations of 0.89 to 0.96 for adults over a seven-year period.
[3] [21] has produced a nice graph to visualize this.
[4] I.e. exponentially sped-up self-improvement – to some extent a strong version of the second requirement.
[5] Unless we connect all pairs of processors together, which then quickly runs into problems of available space as the number of connections is quadratic in the number of processors.

self-improvement by superintelligences, such as corporations like Intel which utilize the collective cognitive resources of tens of thousands of humans as well as millions of older CPUs to design the next CPU generation, and have been doing this for decades. However, this has not led to a hard takeoff; rather, it has led to a very soft takeoff in the form of Moore's law. Naam [21] further points out that the computational complexity to create higher intelligence may be much greater than linear.

It may be argued that evolution on earth has been running just such an experiment at recursive self-improvement for the last 3.5 billion years and has been relatively successful, even creating several human-level intelligences from scratch, all but one of which have already died out.[6] From this we can infer that recursive self-improvement is possible over sufficiently long time spans at the level of species – not necessarily at individual level. In fact the process used by evolution, *Survival of the Fittest*[25], is not reliant on individual intelligence at all. Taking a close look at the many organisms (*solutions*) it is clear that most of them are in an evolutionary dead-end and further optimization is likely to be increasingly difficult (see e.g. [26], pp.104–105).

But although the biosphere contains approximately 16 million different species and $5 * 10^{30}$ living cells and has been "running" for about 3.5 billion years, optimizing all species in parallel, human-level intelligence appeared only relatively recently around 6 million years ago (i.e. after 99.83% of elapsed "runtime"). From this we may infer that creating the higher levels of intelligence so essential for the singularity is exceedingly hard, and generating them from scratch may be very costly.

It may also be argued that evolution does not optimize for intelligence as it may not be a survival trait.[7] However the one long-standing result compatible with this view – that highly intelligent people have fewer children than less intelligent ones – has been reevaluated by Parker [28] who found no such difference after reanalyzing earlier work, and who speculated that sibling density – leading to less individual educational attention and less monetary and educational resources – might also explain these differences. Again we note that intelligence cannot be *improved* over the level obtained by nature and early nurture influences and is remarkably stable in adults, however many effects (nutritional, emotional, education, ..) can *stunt* intelligence below the level that could be obtained under optimal conditions, so these results do not contradict the ones mentioned earlier on intelligence being practically not improvable [20].

It is also sometimes claimed that the number of different ways to the singularity (such as Artificial Intelligence, Intelligence Amplification, hive minds, and so on) increases the likelihood of it to happen. However, this is a rather simplistic view. Notably, if only one of above requirements cannot be met, the singularity will never happen.

Concluding, we find all three mentioned conditions to be implausible. Recursive self-improvement is empirically only shown at species or population level, and may not be feasible on the level of individual intelligences at all, as human intelligence – the

[6] Unfortunately the extinct homonids seem to have been the more peaceful ones, e.g. Kwang Hyun [24] states in his conclusion that the genetic basis for aggressiveness and hyperactivity originate with humans, and human interbreeding with Neanderthals led to more peaceful behavior in their hybrid offspring. This is likely the reason why they are extinct.

[7] Very entertaining argued for in [27] – but please consider last paragraph of Section 1 on the human weakness for compelling stories.

only known working example – is practically not improvable. There are also indications from evolutionary biology that recursive-self improvement is very costly. Finally, the additional exponential process needed to actually get to arbitrarily high intelligence levels – Moore's law is very often used here – merits a closer look as it is a central point of quite a few singularity proponents.

3 MOORE'S LAW REVISITED

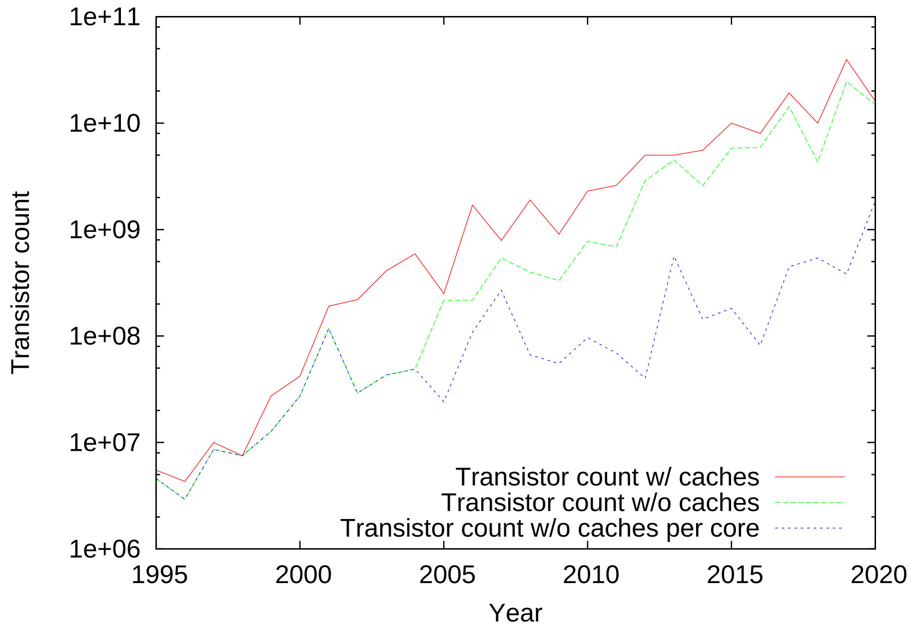
Moore's law has often been cited and used in projections as a support for the impending technological singularity. Even when it is not directly referred to, Grace [29] notes that software costs roughly track hardware costs in many domains, so estimates on such progress – even when not directly referencing Moore's law – still indirectly rely on its exponential improvements. Tuomi [30] makes a very good case that the increasing number of transistors on chips confounds two different patterns; we added a third somewhat obvious observation that may not be common knowledge.

1. In 1965, initially Moore's law – then a empirically grounded hypothesis on the future of semiconductor production technology – referenced minimum-cost chips, i.e. those that could be most efficiently produced. However, ten years later in 1975 it was changed to reference maximum-complexity chips, i.e. at the limit of the production technology, which changes both timing (higher values will tend to be reported earlier) and the transistor count itself (which will also be higher, perhaps beyond profitability).[8]
2. From about 1998, the chip industry started to put cache memory on its chips to improve performance. This increases the transistor count dramatically with only modest increases in design complexity, as cache memory units are relatively uniform and simple compared to other CPU units. This explanation also applies to some extent to graphical processing units (GPUs) used extensively in Deep Learning. For example, the last DEC Alpha chip had 90% of the transistors corresponding to cache memory [31] and the Intel DualCore Itanium-2 from 2006 had 96.28% cache memory [32].
3. From about 2005, semiconductor manufacturers started to fill available die area with multiple copies of chip units (*cores*) that are essentially identical, thus increasing transistor count with only small increases in design complexity.

We cannot easily verify 1) but it was only at the beginning of Moore's law and is no longer relevant. However, we can with some effort confirm 2) and 3) by closely analyzing the processors with the highest transistor counts in each year from 1995 to 2020 (data from [31]). We manually determined transistor count, cache size and number of cores for each processor. To reduce effort, we only analyzed one processor per year, namely the one with the highest transistor count. Fig. 1 shows uncorrected transistor

[8] Also, maximum-complexity chips sometimes had bad price/performance ratio and were hard to sell, occasionally never coming to market at all.

Fig. 1. Transistor counts between 1995 and 2020

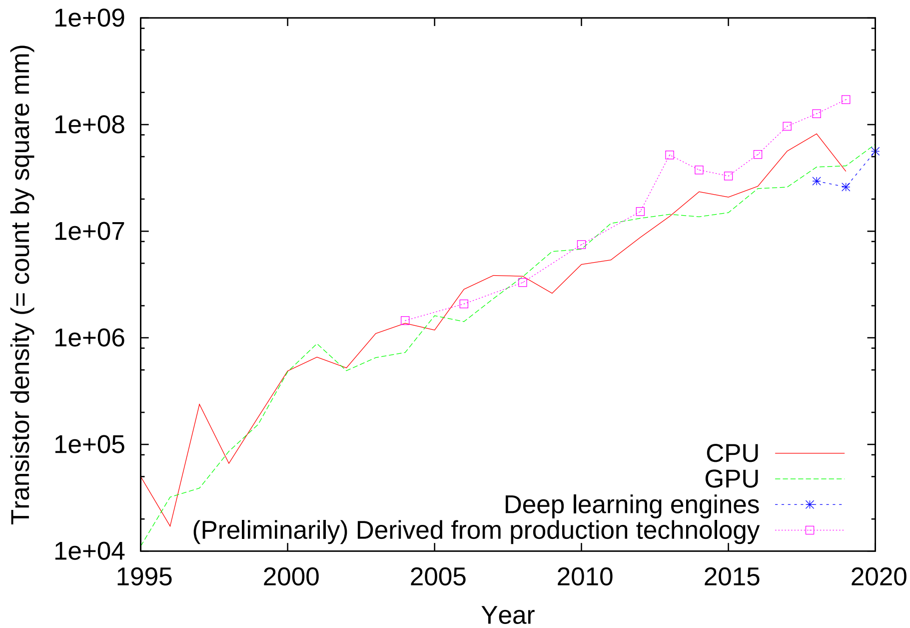


count, transistor count when removing approximate cache transistors[9] (compensating for effect 2), and the transistor count per core (compensating for effects 2 and 3). As can be seen, cache contributed a large proportion of transistors between 2000 and 2005, and its contribution has been shrinking since then. The number of transistors by core shows only a very weak exponential growth and may even become asymptotic in the near future. The doubling time for by-core transistor counts is already 3.84 years, almost twice the doubling time for uncorrected Moore's law (of two years). The doubling time for 1995-2007 in our period is 2.30 years, and for 2008-2020 this has increased to 2.97 years, so we see a slowdown in the by-core transistor counts. This is even more so for the uncorrected transistor counts: here 1995-2007 has a doubling time of 1.36 years – still near the Moore estimate[10] – and 2008-2020 has a doubling time of 2.79 years, almost twice that of the earlier interval.

[9] We estimated seven transistors per bit, which was current in 1999 but may be down to four transistors at present. However on an exponential scale a factor of two corresponds to a constant shift and does not effect the rest of this analysis. Also our method ignores cache logic and thus slightly underestimates the number of cache transistors, which can also be seen from [32] (entry from 2006) where cache transistors are estimated at 1.5 billion by our method, but are actually 1.656 billion out of 1.72 billion (96.28%). We had to use this ad-hoc method as for most processors the number of cache transistors was never reported while cache memory sizes were readily available.

[10] Which is sometimes quoted – and used for long-term planning in the chip industry – as 1.5 years or 18 months.

Fig. 2. Transistor density ($\frac{transistor\_count}{area\_in\_mm^2}$) between 1995 and 2020



We can perhaps infer that maximum design complexity, i.e. the most complex single processor core that can be designed by human minds, will be soon reached.[11] But even the total transistor count shows a similar slow-down of factor two which may indicate an upcoming exponential growth limit.[12]

One alternative interpretation of Moore's law is in terms of transistor density (i.e. $\frac{transistor\_count}{area\_in\_mm^2}$). In Fig 2 we see transistor densities from 1995 to 2020 for CPUs, GPUs, a few deep learning engines[13] and some estimated densities derived from production technologies at the time they were introduced.[14] For 2020, we are already at 5nm technology, but we do not have data on empirical densities. Still, since a silicon atom has a diameter of about 0.5nm we are only a factor of 10 away from the physical limit of this technology. And since in 2007 – about 13 years ago – we had 45nm technology which is roughly ten times larger than 5nm, this means just another 13 years to

[11] It may be argued that it has already been reached according to recently found bugs in almost all modern processors, see [22].

[12] GPUs contain millions of similar relatively simple cores and their transistor count per core would thus be much smaller in this graph, outside the shown range (data not shown).

[13] Only three datapoints.

[14] These are not intended as maximum *achievable* densities, as the technology will – if possible – be improved over time and may give slightly higher densities later, and these densities are also not known for many actual production technologies – only Intel reports some results. So please do not be confused by the seemingly impossible results in 2004, 2006, 2008 and 2010.

go before we reach physical limits, albeit in the unlikely case the doubling time does not increase further.

It also can be seen that CPUs and GPUs follow roughly the same curve, and are thus limited by the same constraints. From this graph we can again estimate doubling times on CPU data[15]: from 1995 to 2007 the doubling time is 1.78 years, and from 2008 to 2019[16] the doubling time was 2.36 years, corresponding to a one magnitude increase in about 8 years. Again the doubling time has increased for the latter interval. So although the absolute theoretical limits are still 8–13 years away, we already see a reduction in the exponential magnitude, where according to singularity proponents we should be seeing a speed-up. Other estimates on when Moore's law will likely stop working are even earlier at 2025 [33, 34], which is consistent with an additional slowdown in the near future.

It has sometimes been mentioned that Moore's law has become a self-fulfilling prophecy, and it is true that semiconductor firms have used it for planning (as doubling in 18 months, see [30]). In fact the reliance on transistor counts for long-term planning is already reduced, amply demonstrated by the fact that Intel stopped reporting transistor counts for their processors in 2017. That there is a challenge to actually fill the technologically available transistor space can also be seen by the fact that from the 2016-2020 top processors, all but one (80%) are a systems-on-a-chip (SoC), and thus include other simple components otherwise part of the mainboard or chipset; and three (60%) include graphical processing units (GPUs) on-chip. This may indicate that we already have too much transistor space to usefully – and profitably – use for mere CPUs.

Additionally, as we noted earlier these observed exponential effects have been obtained by diverse overlapping techniques: from 1975 onward by reporting maximum-complexity rather than minimum cost chips, from 2000 to 2010 by adding relatively simple cache memory, in 2005-2020 by successively increasing the number of identical cores on each die. There is no single underlying process to account for even the by-core transistor counts, so biological metaphors of an uniform growth process fall quite short.

Also, the success of the semiconductor industry has been heavily dependent on external factors, which can be seen by studying the history of the semiconductor industry. This industry faced several challenges, where they were in each case saved by external factors providing a large market for their chips – first for calculators and digital clocks, then for mini- and mainframe-computers, in the mid 1980s for the IBM PC and Microsoft with Personal Computers, and in the 1990s for the World Wide Web, which exploded the hard disk and memory market and also created the need for new powerful processor architectures to handle images, sound and highly compressed video [30]. If any of these events had not happened, Moore's law would have already been broken.

Even worse, as we noted the Singularity requires that costs to achieve this exponential speedup are at most linear in transistor count or density. However, the investment for a new semiconductor factory has also been increasing exponentially – doubling every four years – and reached US$ 14.3 billion in 2015. This is known as Rock's law, or

[15] We did not use GPU data since the number of datapoints was smaller while its shape is very similar to the CPU curve.

[16] No data point existed for 2020 at the time of writing this paper.

Moore's second law [35]. In fact, already in 1979, Moore noted that the man-hours per month required for integrated circuit production were also growing exponentially. That all the effort put into making chips faster using cheap cache memory, pipelining, out-of-order execution, branch prediction, and other methods to increase performance has not been without issues can be seen in the recently found bugs such as Meltdown and Spectre [22] – an indication that we can build things we do not truly understand until decades later and that current developments costs, already quite high, are still underestimated.

3.1 QUANTUM COMPUTING

Quantum Computing (QC [36]) is sometimes mentioned as an alternative to classical computing which may allow higher packing densities. At first glance it looks promising, since it uses *qubits* which can store quantum superpositions instead of 0-or-1 bits. This means a quantum computer with n qubits can store 2^n values and can also run reversible operations on them in parallel. However, the obtained results stored within the qubits cannot be read out directly since each measurement collapses the wave function and forces the qubit's value to either 0 or 1 which depends stochastically on the collapsed superposition. So for an useful algorithm as few measurements as possible must be performed to reconstruct the relevant parts of the collapsed wave function, restarting the quantum computation after each measurement. If this can be done successfully, quantum computers *may* be able to compute some functions exponentially faster than classical computers. Such *quantum supremacy* has however not been conclusively demonstrated and would in any case only be limited to a small set of algorithms proposed for quantum computers. For initializing the qubits, executing operations, and reading out the results classical computers are needed.

Quantum computers can potentially simulate normal computers. However, in this case each qubit corresponds to a classical bit and there is no advantage over classical computing w.r.t. transistor packing density.

It should also be noted that despite the potentially much higher packing density for temporary storage (qubits), the quantum gates themselves can – in almost all currently proposed quantum architectures – not be made smaller than a single atom, limiting the transistor packing density in a similar way as for classical computing.

3.2 SPINTRONICS AND PHOTONICS

A combination of spintronics and photonics could achieve much faster computers running at terahertz frequencies – about 250 times faster than current CPUs – by combining information storage via magnetic spin and purely photonic gates. This is an active research area [37], but no working system with logic gates *and* memory has been proposed as of now. Even if such a system already existed, major additional minaturization would be needed to build a chip competitive to classical designs. If we also managed to shrink logic gates and memory units down to single atoms, it would put us at a point about 30 years beyond Moore's law. However, this would need major breakthroughs in this field over the next ten years and is therefore rather unlikely. Also, IBM [38] has already demonstrated a 100 GHz transistor in 2010, so higher clock rates may also be achievable using more classical transistor designs, negating this speed advantage.

3.3 REVERSIBLE COMPUTATION

Classical computation works irreversibly and has therefore a lower bound on heat dissipation due to entropy increase [39] which will become more relevant the closer we get to single-atom transistors. One possible solution for this is to use reversible computation which does not generate heat. This can be done by using reversible gates. However, reversible computation is inefficient in the sense that many bits must be preserved to enable reversibility which do not contribute to the actual computation and thus reduce the effectively usable transistor packing density. Also, if reversibility cannot be completely obtained, it may be necessary to run a reversible computer very slowly to limit heat dissipation. Such reversible computers are needed to drive quantum computing as even the small heat dissipation from the erasure of a single bit may destroy the fragile quantum states on which the operation of the quantum computer relies on. It is clear that such designs will always be slower than corresponding designs using irreversible computation and appropriate cooling.

4 A CASE FOR HISTORY

In light of these negative results, why is the myth of the technological singularity still so compelling? To better understand this, let's take a look at AI history, starting with the precursor of AI – Cybernetics. This field is similarily contradictory and emotionally burdened as the singularity discussion, and retained its own persistent myth, as this quote shows:

> "One pattern is spiritual. Not always, but often, the machine has become a godhead, an idol... Science created a totem. The machine became the avatar... A second powerful pattern is contradiction... Automated factories would free workers from undignified drudgery, yet deprive them of their dignity... Computers were dumb and could be hacked by teenagers, yet they could outsmart humans... More networked computers would lead into a "dossier society" of ubiquituous surveillance - and enable anonymity and a freer and better political order. Networked information system would make nations more vulnerable and more fragile than ever, and networked command-and-control systems would make their armies more dominant and more lethal than ever. Machines would be future society's hard-charging overlords and its soft underbelly. The myth [of cybernetics] hides these contradictions and makes them acceptable."
> (Rid [40], p.348f, quoted with friendly permission by the author)

The myth of the upcoming technological singularity is very similar in spirit to the myth of cybernetics. In a way, the discussion on the singularity reflects earlier cybernetics discussions now only known to students of AI history.

In the research proposal for the Dartmouth meeting in 1956 – where the term *Artificial Intelligence* was coined – it was proposed that "a 2 month, 10 man study of artificial intelligence be carried out" and it was concluded "We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer." [41]. Now, in 2021 – 65 years later – it is clear that the scope and breadth of the task has been severely underestimated by many orders of magnitude.

But Artificial Intelligence actually starts a bit earlier than that. We can see it already starting in 1943 with McCulloch and Pitts [42], who proposed a boolean circuit model of the brain.[17]

In 1950, Turing [43] reformulated the question of whether machines can think to whether machines can behave intelligently, and proposed an operational, non-constructive and non-reproducible test for intelligence – the Turing test. He suggested major components of AI: knowledge, reasoning, language understanding and learning, and anticipated all major arguments against AI. Nowadays, Chatbots based on text-mining terabytes of chat room logs are often judged intelligent by non-experts [44], however they still have practically no chance to fool an experienced AI researcher.

Early AI successes like Samuel's checkers program [45] which could beat human players and was trained only against itself, Newell and Simon's Logic Theorist [46] who found more elegant proofs for some known mathematical theorems, and Winograd's Blocks world [47], which could answer questions and act in a simplified blocks world with the abilities of a three-year old child have initially confirmed the optimism at the Dartmouth meeting.

However when we look closely with the benefit of hindsight, we note that – for example – checkers has a very simple structure that makes the simplistic approach work where it failed with more complex games such as Chess which was only solved many decades later. Another example is Backgammon, where a temporal-difference learning algorithm [48] achieved grandmaster-level performance with just a handful of heuristics. It was later found that the game's structure fit well with the coevolutionary learning paradigm [49], which explained most of its success and why temporal difference learning did not work on many other games. The blocks world AI actually only understood very specific simple sentence patterns. Even slightly different formulations or typing errors make the system fail to understand – it is extremely brittle like most early AI systems. Now, 65 years later, we are finally able to build systems which understand written and even spoken speech far more robustly, but it has been extremely hard to get here.[18]

One of the reasons why it has been hard was computational complexity. Most hard problems get exponentially harder with larger problem sizes, meaning that small instances (toy problems) can be easily solved, but even slightly larger problems take years, and realistic problems may take longer to solve than the remaining lifetime of the universe.[19]

The solution proposed around 1969-1979 was knowledge-based systems, resulting in the birth of the expert systems industry. However, knowledge as stored and processed by machines is very different from knowledge stored and processed by humans.

[17] See Section 8 for an overview of different brain models that have been proposed in known history.

[18] Compare e.g. the simplicity of TD-Gammon learning with AlphaGo [50] w.r.t algorithmic complexity and the amount of computational power needed to train the system.

[19] According to the big crunch theory. If our universe were to expand indefinitely, we would still at some point run out of energy, but it would be much later.

For example, contradictory statements make a knowledge base completely useless[20], which makes it exponentially harder to build larger knowledge bases, as they have to be completely free of even the smallest contradiction. So this problem – exacerbated by overarching promises – later lead to a bust in the expert systems industry: the so-called *AI Winter*.

From 1988, there was a resurgence of probability, increase in technical depth, and the birth of Nouvelle AI: Artificial Life, Genetic Algorithms, and soft computing such as Fuzzy Logic. From 1995 the beginning of the Agents metaphor, a stronger focus on embodiment – another thing which is coupled strongly with human intelligence – and an increasing number of real-world applications was observed.

The Dot-Com crisis from 2000 reduced investment into computer technology for some time, so in the first decade of the 21st century, not much progress was made. Still, in 2005 Stanford's Stanley [51] won the DARPA Grand Challenge, driving autonomously 131 miles through the desert without using a camera. Laser-range LIDAR sensors were used throughout and almost all parts of the system were trained and tuned using machine learning algorithms, showing a clear advantage over the handcrafted systems used by competitors.

In 2010, Microsoft introduced the Kinect 360° motion sensor, which had been trained using Random Forest classifiers [52] to recognize human figures from depth camera[21] recordings. In 2011, IBM's Watson [53] beat the two greatest Jeopardy! champions.

Deep Learning took off in 2012 with the seminal paper by Hinton et al. [54]. This was due to several factors – wide availability of software frameworks, cheap GPU hardware due to the successful gaming industry, and better learning algorithms – which we will go into more detail later in Section 5. In 2015 Tesla announced a software update to enable self-driving for their electric cars on freeways. Their system was end-to-end trained using Deep Learning, albeit using only cheap 2D cameras and radar, and is therefore not as robust as LIDAR-based self-driving systems. In 2016, Google DeepMind AlphaGo [50] beat 9dan Go champion Lee Sedol 4:1. In 2017 AlphaZero [55] demonstrated learning to play Chess and Shogi (chinese chess) at champion level in days. Rational solvers have been optimized to such an extent that large proofs can be generated in days, such as the proof that the 5th Schur number S(5) equals 160 [56]. This proof is 2 petabytes large, and needed 14 CPU years computation time. In 2019, AlphaStar [57] won in the real-time strategy game Starcraft II 10:1 again human expert players. In 2020, AlphaFold [58] won the biannual Critical Assessment of Protein Structure Prediction competition by predicting the 3D structure of proteins much better than the closest competitor. So it is clear that Deep Learning dominates AI research at present. But how did this happen?

[20] Known as *ex falso quodlibet*: from false [knowledge] anything can be derived.

[21] A depth camera returns pixel arrays similar to a normal camera. However the pixel values correspond to the measured distance of each pixel from the camera plane rather than brightness or color. These distance measurements are usually obtained via active sensing.

5 THE RISE OF DEEP LEARNING

No criticism of the singularity would be complete without a section on Deep Learning. The history of deep learning started with the already mentioned work of McCulloch and Pitts [42] who presented a simplified model of nerve cells, which computed a thresholded step function on a weighted sum of inputs. Rosenblatt [59] introduced a formal definition for such a system, coined the term *perceptron* and demonstrated a learning algorithm that could learn simple input-output functions such as AND, OR and NOT in a single perceptron.

Minsky and Papert [60] incidentially proved in 1969 that a single perceptron could not learn the XOR function – or any other not linear separable function. Although Rosenblatt could demonstrate – as McCulloch and Pitts [42] had done before – that a combination of several perceptrons could learn XOR, the lack of a suitable training algorithm for such combinations of perceptrons still lead to an almost complete stop in research on neural networks for about a decade. The perceptron remains in its original and modified forms the core unit for neural networks, including convolutional and deep learning networks.

Only when Werbos [61] independently reinvented concepts already known by Kelley [62][22], and proposed the backpropagation learning algorithm for training networks of interconnected perceptrons arranged in layers (multi-layer perceptrons, MLP[23]), research continued albeit at a relatively low level. Rumelhart [64] soon demonstrated that this learning algorithm led to useful internal representations for the internal layers.

LeCun [65] introduced a new form of neural network which was inspired by the visual system[24] and arranged the perceptrons in their natural 1D, 2D or 3D order and introduced two new types of layers:[25]

- Pooled or subsampling layers, which compute global statistics, such as maximum or average over a window of perceptrons from the previous layer, just like a simple filter.
- Convolutional layers, where weights are shared between perceptron units, i.e. each point of the next layer computes a sum with the same weights, but on a different window of the previous layer – a filter or convolution in image space.

Such a network was trained to recognize handwritten digits for the US postal service and similar networks are still used by many OCR systems. Lastly the seminal paper by Hinton et al. [54] to which all the field's major research labs and researchers contributed

[22] Tragically, this work would already have been available when Minskey & Paperts book was published. However since it was a completely different application and research field, this was not noticed until much later.

[23] A second way to solve the XOR-problem for perceptrons is to regularize the model space by means of maximum margin hyperplanes (also making the solution unique) and introducing the kernel trick – expanding the input feature vector with base functions. This lead the the field of Support Vector Machines (SVM, [63])

[24] It however ignores the 90% LGN feedback connections which we will mention in Section 8.

[25] Previously, only fully connected layers were used where each perceptron within one layer is connected to all perceptrons within the next layer.

lead to renewed interest in this field and – due to the new focus on networks with many levels, although it also introduced new layer and interconnection types and perceptron activation functions – it was renamed to Deep Learning.

So why did it take Deep Learning so long to take off? Earlier, it was generally believed that neural networks were very unstable learning systems and not suited for a wide range of problems. It was believed that overfitting and getting stuck in local minima would be a major problem for many tasks. Also, it was considered infeasible to obtain feature sets competitive to handcrafted features, especially for image and audio processing which had been researched very intensively up to this point. The lack of publicly available source code to train such systems was also a problem. For example, we replicated work by LeCun [65] in [66], but the source code used to train the convolutional network model was only available under a cumbersome license, and heavily tailored for the task of handwritten digit recognition, and thus could not be used for other projects.

To enumerate more systematically, the main reasons why Deep Learning took off were the following:

1. Moore's law and the corresponding increase in single and multi-processor performance since 1970. This was especially important to Deep Learning, as classical (human-written) machine learning algorithms are very seldom highly scalable[26] – i.e. they profit linearly from *faster* processors, but much less from *more* processors. Now that Moore's law is mostly driven by integrating even more parallel processors, classical code profits much less from the speed-up than Deep Learning. In fact, Deep Learning training algorithms are almost infinitely scalable and can be run in parallel with one processor per perceptron unit, of which there can be several million, albeit communication costs will at a certain point begin to dominate, limiting scalability.
2. The resurgence of the vector processor in form of graphical processing units (GPUs).[27] This factor was driven by the gaming industry in its goal to create ever better computer game graphics on similarily improving monitors (in terms of resolution and update frequency). Training and evaluation of Deep Learning networks basically involves computing relatively simple, very similar operations on a large set of units. A normal CPU can of course perform these operations, but is optimized for a more diverse workload and normally slower than a specialized GPU for such simple essentially vector operations. It turned out that GPUs can easily be used for Deep Learning and initially speeded up training by one – and later two to three – orders of magnitude, and in fact there are now specialized processors optimized for Deep Learning tasks that obtain even higher speedups.
3. Once a large set of research groups applied these methods, it became obvious that the perceived disadvantages (instability, local minima) were not significant and could be overcome relatively easily. In fact Deep Learning systems soon dominated visual recognition, image classification and speech recognition tasks within a

[26] Except perhaps for Random Forest [67], which we've applied in unpublished research and found to scale very well.

[27] If you remember the large Cray computers – yes, these were vector processors as well and they could have been used for Deep Learning already in the 80ies, forty years ago as of now.

few years, thus confirming that they are able to create internal feature spaces (represented within the first few layers after the input layer) that are superior to handcrafted human feature spaces in many tasks. This does not mean that Deep Learning system are without issues – failure modes as described in [68][28] and [69][29] indicate that their models are quite different and less robust than the ones human brains use.

4. New and improved training algorithms such as Adam [70] reduced both training time and the risk of getting stuck in local minima, which had previously been a significant problem in training deep learning networks. Initializing the model weights not randomly but with statistical methods (e.g. PCA) also adressed these issues and furthermore speeded up training time significantly. Both made it possible to train larger and more complex network models.
5. With Torch (2002), Theano (2007) and especially Tensorflow (2015) [71], a large set of deep learning frameworks with GPU support became mature and widely used. Although Theano has ceased development in 2017, and Torch in 2018, pyTorch – a fork of Torch written in Python – is still in active development as of December 2020.
   Especially Tensorflow, which contains a large set of samples and pretrained models including training code and in most cases the data needed to retrain the model itself – combined with the interface in the easy-to-learn programming language Python – has made Tensorflow very popular. From version 1.5, models could be converted to Tensorflow Lite to run on mobile phones and embedded platforms such as the Raspberry Pi. All these frameworks – most of which are available under a permissive license such as MIT or Apache 2.0 – made applying the corresponding learning algorithms much simpler, and led to a corresponding increase in researchers using deep learning, and therefore also in the number of applications.

But does all this mean we have already – or will shortly have – achieved human-level intelligence, i.e. an Artificial General Intelligence (AGI)? Sadly, no.

6 THE FALL OF DEEP LEARNING

One major reason why Deep Learning is not an Artificial General Intelligence (i.e. a human-level AI) is that our training algorithms create models that have completely different – and far more brittle – failure modes than human intelligence, amply demonstrated by systems such as DeepFool and ColorFool [69, 68]. So the learned models are quite different. This alone however could simply be an artefact of the network architecture. However, it is also quite inconceivable that even relatively simple learning algorithms such as Backpropagation – to say nothing of Adam – could be actually used within our brains. The learning algorithms are much too different. Also, the nerve system model of Deep Learning still basically reflects the perceptron and is by no means a biologically plausible model as this incomplete list of open issues shows.

[28] Arbitrarily changing object classifications and confidences by specific changes in background colors for input images.

[29] Arbitrarily changing object classifications and confidences by adding specific noise to the input image.

1. It is unclear how the goal values could be transported through all neural layers to compute the error feedback. Also, despite decades of experiments, no backpropagation error signal was ever observed.
2. Biological neurons communicate via binary spikes and not via numerical values
3. Synchronization between neurons – an extremely important mechanism used in the brain – is not modelled in almost all Deep Learning networks.
4. Backpropagation would need perfectly synchronized discrete steps which is not possible as biological nervous systems do not have a sufficient precise global clock.
5. The weights for the backpropagation connections would have to mirror the weights of the corresponding forward connections. There is no known biological mechanism which could ensure this.
6. Backpropagation would need to know the non-linear derivates of all neurons in the higher layers. Again, there is no known biological mechanisms which could – even potentially – compute these values.

It is clear that Deep Learning reflects one deeper principle on how neural networks build complex models, but it is also similarily clear that it is just a small part of the whole picture.

Additionally, the original inspiration for convolutional neural networks came from the human visual system – however the 90% feedback connection to the LGN nucleus from almost all other brain areas were completely ignored. But such massive feedback loops may be essential for robustness (more details see Section 8).

The DeepMind series of AlphaGo [50] and AlphaZero [55] models is perhaps best suited to show the disadvantages of Deep Learning from a different view point: First, these system can only be trained by self-playing resp. by creating extremely large data sets (sometimes implicitly) as millions to billions of samples are needed. No human player can generate even a small part of the training data these systems need to get up to speed. This also means that if one player notes a way to beat the system – easier in perfect information games such as Go, Chess or Shogi, since there are no random fluctuations in the environment – it is always possible to beat the system in exactly the same way, and there is no easy fix for this.[30] It's possible to force the system by re-training to choose another path at a critical juncture, however this might weaken overall performance, and in any case there will always be other weaknesses to exploit. What cannot be done is to train the system to play well against a specific player since – as noted above – even if this player played Go games their whole life, and all games were stored and trained, this would not even make the smallest difference.[31] Especially for AlphaStar [57] (online first-person gaming) it was mentioned that initially the model used recorded human player movements as otherwise the exploration state space would simply have been too large to find meaningful movement patterns. However note that children and teenagers have no problems learning to play this game without such information in single player mode.

[30] We may assume the creators of AlphaGo are aware of this issue and perhaps it informed their decision not to make AlphaGo publicly available.

[31] This was clearly mentioned at one of the press conferences after the AlphaGo win, and should of course be obvious considering the large amount of data such systems need for training.

Realistic estimates towards when Artificial General intelligence will be available tend towards several centuries. As mentioned Hanson [17] (p.63) estimated 4-8 centuries should Moore's law reduce just by a factor of two but still keep exponential over the whole time period, which is actually very unlikely as we will hit strong physical limits in the next decade. Since the beginning of AI research in 1956, the median forecast of the duration of time until human-level abilities would be achieved in AI research was always around 30 years. Needless to say, all these forecasts were wrong.

So, sadly, although Deep Learning is a big step towards understanding intelligence – or more precisely, how extremely simplified pseudo-nervous system can be trained using biologically implausible learning algorithms to solve tasks that for humans require intelligence – but not an AGI at all. Many further significant advances would clearly be necessary to achieve this.

7 SINGULARITY CONTRADICTIONS

Should we somehow still manage – by serendipity, sheer luck or perhaps as a download from an extinct extraterrestrial civilization – to create an Artificial General Intelligence, it is by no means assured that it will achieve superintelligence. There are quite a few somewhat obvious contradictions within the story of the technological singularity which would also apply to any AGI (including ourselves, of course).

7.1 THE MYTH OF BETTER INTROSPECTION

It is seldom questioned that an AGI would be able to bootstrap itself to higher intelligence quite easily. However, why would an AI at the same intelligence level as us have better insight into its internal structure? Why should its mind work parallel and not serially as ours, limiting the obtainable speedup as well as self-reflection? We note that although the human brain is a massively parallel system, the human mind is still only able to do one thing at a time and is very bad at multitasking. It is not obvious why such a constraint should not be a general constraint of all minds, perhaps driven by the necessity for limited self-awareness and an unified stream of consciousness.

A superintelligence is expected to be able to fully understand humans. But how to get from a human-level Artificial General Intelligence to a superintelligence when it cannot improve itself much better than a human? The exponentially increasing speed of self-improvement, so essential for the technological singularity, is not likely to hold at any intelligence level. Otherwise programming code would not still be written by humans. It is for example possible to train Deep Learning networks with a few millions lines of C code. They will then output syntactically correct code with few errors (mostly concerning assignments to previously undeclared variables) that is completely meaningless. Without a guiding mind and intentionality behind it, even extremely subtle abilities are essentially useless. And a true AGI written e.g. in C that would be as proficient as the best human C programmer would still be no better suited to rewrite itself to become smarter – similar as we humans do not have a privileged insight into the workings of the 10 billion neurons of our brain.

Even if a superhuman intelligence could be created somehow, it would potentially be able to improve lesser human-level AI systems – perhaps even to its own level – but it is rather unlikely that it would be able to indefinitely improve itself at ever increasing rate. As any programmer knows, radical speedups can sometimes be obtained, but the law of diminishing returns sets in and at some point the effort to make a program go faster is simply not worth the obtainable speed-up. In many cases, optimizations are even detrimental and reduce performance. It is unclear why this should be different for the extremely complex operations normally associated with high-level cognition. If anything, one would expect even steeper diminishing returns here. And improvement in hardware is time-consuming and costly. A new semiconductor plant takes billions of dollars to set up and only runs cost-effectively after several years of extensive experimentation and optimization. To believe a superhuman intelligence can cut through all these issues, and ignore physical reality and all manufacturing constraints in building its successor is only a pipe dream. Especially since it is well-known that building robots that are as flexible and versatile as the human body is extremely hard, as anyone who has ever built robots can attest to. Despite decades (soon to be a century) of research the state-of-the-art bipedal robot apocalypse could at present be fended off at the top of stairs with slingshots. In fact embodied robots instantly seem much more intelligent than their simulated counterparts even if they are not, simply by utilizing the randomness of real environments, and it is unclear whether an intelligence could ever be created without some kind of grounding in physical reality – which would again preclude superfast self-improvement.

Concluding, we find it increasingly implausible that even a superhuman mind could bootstrap itself to a significantly higher intelligence level, and more so that it could do so at an ever increasing rate (as in Kurzweil's [12] law of accelerating returns). So even if a superhuman AI would magically drop into our hands, it would not cause a singularity. And it would have to drop, since despite 65 years of research we have actually no idea how to build one, or even just an AGI.

7.2 THE MYTH OF CONTINUED EXPONENTIAL ECONOMY GROWTH

Assuming an economic doubling time of 15 years (assumed by [17]), a century means about seven doublings, meaning that even in one century the economy would be $2^7 = 128$ times larger than it is now. As we mentioned, AGIs – surely a necessary precursor of the singularity – are at least two and more likely 4-8 centuries away. Historically, there has been a strong correlation between economical growth and energy consumption, but even a small poportion of a 128-fold increase in energy consumption would probably change climate to such an extent that human survival is not longer feasible.

The singularity story mainly counters this with *superexponential dematerialisation*, i.e. the theory that it is possible to obtain similar growth rates while using ever less energy and matter. This implies not only an uncoupling of energy consumption and growth, but an inverse correlation – something that was never seen in the past of the industrial revolution for the last 200 years.[32] But as the necessary technologies are at

[32] It is also quite unlikely that it ever happened before in human history, seeing that for most of history almost all people lived near subsistence levels.

least several decades in the future – if they can be made to work at all – this will come too late as only a single doubling in energy consumption (and thus CO^2 emissions) will be enough to change the biosphere in a way that makes living on this planet no longer feasible for the number of people who are currently living here. Much more likely is a radical reduction of growth rates in energy consumption: for example Sharma [72] quotes a global growth rate of energy of 0.8% for 2016-2030 and 0.1% for 2030-2060. This would mean diverting time into more efficient processors which need less energy, and reducing the expected growth rate in computer performance further.

For comparison, the human brain consumes about 50W of energy and generates this energy in a completely sustainable way, while Google consumed 10.6 Terawatt hours in 2018, which is up from 2.86 Terawatt hours in 2011, corresponding to a growth of 20.58% per years, or a doubling time of 1.66 years (20 months), even faster than the doubling time in transistor count. It is clear this cannot be sustained for even one additional century.[33]

In fact Gordon [73] points out that measured economic growth has slowed around 1970 and slowed even further since the financial crisis of 2007/2008, and also argues that the economic data show no trace of a coming Singularity as imagined e.g. by Good [5] already in 1966. So the growth observed due to Moore's law seems to be the exception rather than the rule, but even this is slowing down as we already showed.

7.3 THE MYTH OF RECURSIVE SELF-IMPROVEMENT IN RESEARCH

Another reason why the arbitrary speedup so desparately needed for the technological singularity is a mirage: The reason why research follows phases is that first a critical set of observations must be present before a theory can be formulated. This process cannot be arbitrarily sped up – e.g. new physical measurement devices must be designed, tested and applied. For example, in physics before Quantum Mechanics, the consensus in the field was that basically most of physics was solved, and only a few strange quirky observations remain which would soon be resolved – such as black body radiation, which is predicted by classical theory to be infinitely high when integrated over all wavelengths. This was called the Rayleigh-Jeans law or ultraviolet catastrophe. Quantum mechanisms quantized the photon output and thus obtained results consistent with observations, but most of physics had to be reevaluated in light of these new assumptions.

In fact Alston [15] found that research funding gives far less than linear returns. It is inconceivable that any kind of intelligence is somehow completely immune from these fallacies and pitfalls, especially since the requirement for recursive ever-faster self-improvement prevents its grounding in physical reality, which may be an essential condition for intentionality, consciousness and human-level intelligence.

[33] Google claims to use 0.3 Wh for the average query. With this amount of energy an average human can think for 21.6s, which *might* just be enough to remember what you were searching for yourself instead of googling it.

8 MIND IS ALWAYS THE HIGHEST AVAILABLE TECHNOLOGY... AND ALWAYS WILL BE.

It is part of a long-standing tradition in human history to take the newest technology and propose that humans – or the human mind – work that way. For example, Galen of Pergamon (129–216 CE) thought that the psychic pneuma (cerebrospinal fluid) who circulate in the ventricles (liquid-filled holes in and around the brain) is conducted along the nerves (thought to be hollow) and effects all psychic phenoma such as sensation, elaborating thoughts and movements – i.e. the mind [74].[34] This was at a time when the first distillation experiments were conducted, and parlour tricks such as boiling wine and igniting the alcohol vapours inspired the idea that invisible influences may be present in liquids.

In 19th century, Victorian England, Huxley [75] proposed that human react as automata in all our apparently free decisions. This was probably inspired by ingenious automata such as [76] – an automaton that could write three different poems and four different drawings; [77] – a mechanical duck that could quack, muddle the water with its bill, drink and eat and seemingly defecate (although the last part was faked); and [78] – an automaton that could reproduce natural speech, as well as countless others including numerous fakes such as the Mechanical Turk. The astonishing performance of these devices led to Radical Behaviourism where the metaphor of humans as automata was upheld for decades despite ample evidence to the contrary.

We already mentioned McCulloch and Pitts [42] who proposed a boolean circuit model of the human brain at the beginning of the computer era, where boolean circuits were the highest available technology.

Quantum mechanics is the theory that brought us – among other things – extremely fast computers, lasers, DVD and bluray players, LEDs and the atomic bomb – and can thus also be considered the highest technology a few decades later. Therefore it should not surprise us that Penrose [79] proposed a quantum-mechanical mechanism to explain the mind, which has been taken up by many other researchers (for an overview see Adams [80]) – although admittedly many of their hypotheses cannot be falsified.

In the 21st century, we now have Deep Learning and again it is assumed that the brain works in a similar way. However, this is a fallacy for several reasons. For one, the best networks are trained used ground truth data in a teacher/pupil setting, while most of the complex work of the brain (such as learning head/eye coordination, segmentation of world images, walking, learning to speak, exploration of unknown environments, ...) is done with no or almost no explicit external feedback. We simply do not have unsupervised learning algorithms that work as well as the supervised learning algorithms commonly used. Secondly, trained networks can be easily provoked to make mistakes, either by adding specific noise into the input images [69] or by specific changes in background colors [68]. Human image segmentation is not susceptible to either attack, in practice the changes – especially of the latter – are so subtle that they are hardly visible even when the original and modified images are shown side-by-side, but they still

[34] Others such as Aristotle saw the brain as mechanism intended for cooling blood, and instead the heart as the seat of soul and mind.

completely confuse DL systems. Optical illusions in humans are on the other hand far more subtle and complex despite our eyes recording images of much worse quality.

A close look at the brain wiring might give us an answer as to why Deep Learning networks modelled on the visual system have these massive vulnerabilities: the lateral geniculate nucleus (LGN), where 90% of the retinal axons (who carry information from the eye) terminate, gets only 10-20% of its input from these connections ([81], p.532). Surprisingly, 80-90% of its input connections originate in many other regions of the brain.[35] It is quite likely that these other connections modulate and adapt image processing yielding the practically perfect segmentation we effortlessly create based on rather bad "eye camera" input. However, if that is the case, to get similar performance, we must also emulate the rest of the brain – the task of image segmentation is then AI-complete, and thus needs an Artificial General Intelligence (AGI). Again, reasonable estimates place AGI at least two, and more likely 4-8 centuries into the future ([17] p.61–63).

9 CONCLUSION, DISCUSSION AND OUTLOOK

We are not overly concerned that the exponential growth in computing performance will quite likely stop in the near future, as it will also take care of planned and unplanned obsolescence, the huge amount of electronics garbage generated by neglecting the actual possible lifetime of electronics devices, the abysmal conditions under which people prospect essential heavy metals such as lithium – the main component in rechargable batteries which are major parts of smartphones, notebooks, and electric cars – and the large amount of time we spend to reacquaint ourselves with new versions of software which are often less intuitive and slower than the old versions.

On the downside we will have to think more deeply about optimizing algorithms and creating – or training – newer, more efficient ones. But since we already have more computing power than we can sensibly put to use, that is not a large issue. We can no longer expect our old systems and algorithms to get exponentially faster automatically. But we won't have to buy or rent new hardware every few years to realize these gains, which more than offsets these disadvantages.

So where do we go from here? We have taken the liberty of writing up the areas where – in our humble opinion – too little is done presently.

1. Currently, Deep Learning algorithms focus mainly on the supervised learning case. However, unsupervised and reinforcement learning in realistic settings – i.e. much less training data, transductive and transfer learning, and especially embodied learning on robots interacting with the physical world – is severely underrepresented. It is our conviction that simulations and offline-learning, albeit useful, can never replace the interaction with physical reality, and that this is the one area where Artificial General Intelligence (AGI) will – eventually – emerge (in 4-8 centuries).

[35] Confirmed are connections to the visual cortex, superior colliculus, pretectum, thalamic reticular nuclei, local LGN interneurons, mesencephalic reticular formation, dorsal raphe nucleus, periaqueuctal grey matter, locus coeruleus and the optic tectum (superior colliculus).

We would go so far as to propose that each AI researcher should be forced to build robots as part of his/her education.[36]

2. We have many frameworks for Deep Learning on sufficiently powerful computers. However, there have been only a handful of systems that allow learning directly on the robot where power consumption, runtime and computational power are severely restricted. We do not yet have the learning algorithms that are sufficiently fast to learn big networks in real-time on these platforms, but perhaps we can work on this. Theoretically, streaming data to a larger platform via Wifi can resolve this to some extent, however our experiences were mixed since latency is a large issue and cannot be completely controlled, also the transfer of uncompressed video frames[37] is more severely limited by latency than bandwidth [83], so the upcoming 5G standard which focusses mostly on bandwidth will not be much help.
3. Deep Learning algorithms are not at all implemented in a biologically plausible way. Although good results are often obtained, it is unclear how our learning algorithms could ever be implemented in an actual biological neural system. So it is quite clear than biological systems must use different – and clearly superior – learning algorithms. There is much work on how biological in vitro neural assemblies can be trained for tasks, but again the underlying algorithms are not known. For example, synchronization between neural assemblies – an important mechanisms for e.g. binding and attention control – is almost completely ignored in the Deep Learning community. It would be interesting to study the actual algorithms used by biological neural systems which after all have had much time to perfect their methods.
   A small model organisms, C. elegans, commonly used for ageing research, has a very small nervous system of only 302 nerve cells. Although its nerve cells are different from mammalian nerve cells[38], if we could record sufficiently detailed traces of the complete nervous system, we may be able to decipher its workings. This organism has six known high-level learning behaviours, some of which can also be triggered in the adult specimen. It would therefore be a reasonable starting point for such an analysis, perhaps better suited than large biological neural assemblies as it is a fully embodied and biologically grounded organism. We have done some minor work on this organism in [84], but so far nobody has proposed a recording system that can reliably record all neurons of this organisms with sufficient temporal and spatial accuracy. Once this is solved, progress may be rapid.
4. The human brain uses about 50W of energy. However the capabilities of the human brain are so much higher than the best currently available platforms which used comparably amounts of power. There are several recent successes in AI e.g.

[36] Of course this will not always work – some roboticists sometimes get caught in these myths as well, e.g. [10]. So perhaps a psychological course on the dangers of antropomorphisation should also be mandatory.

[37] We note that the existing compression algorithms have been optimized to remove parts of images and videos not observable by the human eye and are likely suboptimal for a non-human visual system. At the very least, training should be done on exactly the same data which was originally collected as models trained on compressed and uncompressed data differ and can not be used interchangably [82].

[38] They do not have Ca_{2+} channels and therefore do not exhibit action potentials.

AlphaGo [50], but the computers and GPUs needed to achieve these successes use approximately 6-10 orders of magnitude more power than the human brain. If nothing else, it would be interesting to focus more resources on researching energy-efficient computing systems and platforms. To some extent this has already begun. This point somewhat overlaps with 2.
5. All minds need self-representation, and it is quite unclear how even modest self-improvement at the highest level can happen without such a self-loop representation. However, presently most Deep Learning networks are either not recurrent (i.e. the do not contain loops) or they only loop inputs or single intermediate layers, not exhibiting the rich kinds of loops between many areas apparent in the human brain or other nervous systems. The reason for this is mainly limitations of our training algorithms and lack of suitable training data (see also point 1 on the necessity for more research into unsupervised learning). Hofstadter [85] has also proposed such loops as central to mind and intelligence.

These are the things we can do as small steps towards an AGI. However, serendipity may achieve what we cannot directly aim for. While we do not believe AGI will emerge spontaneously on the internet (as imagined in [86]) – the internet, although rich in content and structure, is neither as complex as the real world nor does it have a sufficient set of strict rules that govern possible interactions and serve as seeds for symbolic computation – it could perhaps emerge in a given sufficiently complex computational structure that also includes interaction with physical reality (as imagined by Hogan [87]). The same author has also written a very insightful critical book on Artificial Intelligence [88] which – although somewhat dated – still contains many valid criticisms.

But let's not cite only science fiction authors on this: Austrian Nobel Prize winner and physicist Erwin Schrödinger notes in the epilogue of *What is Life?* and also to some extent implies in *Mind and Matter* that all minds must be implemented at the lowest level of physical reality, i.e. "I ... am the person, if any, who controls the 'motion of the atoms' according to the Laws of Nature." ([26], p.87). A candidate mechanism, *Orchestrated Objective Reduction*, has been proposed by [79] and defended in [89], however it is quite implausible and has been severely critized. Still, if we follow Schrödinger on this, mind must be found at quantum levels or below. So perhaps we will someday obtain valuable information from the unlikely direction of fundamental physics research as well.

References

1. Vinge, V.: Marooned in realtime. Macmillan (2004)
2. De Abaitua, M.: The Red Men. Hachette UK (2013)
3. De Abaitua, M.: The Destructives, vol. 3. Watkins Media Limited (2016)
4. Ulam, S.: Tribute to John von Neumann. Bulletin of the American mathematical society **64**(3), 1–49 (1958)
5. Good, I.J.: Speculations concerning the first ultraintelligent machine. In: Advances in computers, vol. 6, pp. 31–88. Elsevier (1966)
6. Lem, S.: Golem XIV. Krakow: Wydawnictwo Literackie (1981)
7. Vinge, V.: First word. Omni **10** (1983)

8. Vinge, V.: Technological singularity. In: VISION-21 Symposium sponsored by NASA Lewis Research Center and the Ohio Aerospace Institute, pp. 30–31 (1993)
9. Solomonoff, R.J.: The time scale of artificial intelligence: Reflections on social effects. Human Systems Management **5**(2), 149–153 (1985)
10. Moravec, H.: Mind Children. Harvard University Press, Cambridge, M.A. (1988)
11. Kurzweil, R., Richter, R., Kurzweil, R., Schneider, M.L.: The age of intelligent machines, vol. 579. MIT press Cambridge (1990)
12. Kurzweil, R.: The singularity is near: When humans transcend biology. Penguin (2005)
13. Yudkowsky, E.: Three major singularity schools. Singularity Institute Blog. September **30** (2007)
14. Sandberg, A.: An overview of models of technological singularity. The Transhumanist Reader: Classical and Contemporary Essays on the Science, Technology, and Philosophy of the Human Future pp. 376–394 (2013)
15. Alston, J.M., Andersen, M.A., James, J.S., Pardey, P.G.: The economic returns to US public agricultural research. American Journal of Agricultural Economics **93**(5), 1257–1277 (2011)
16. Hanson, R.: AI Progress Estimate. Overcoming Bias blog, August 27 (2012). URL https://www.overcomingbias.com/2012/08/ai-progress-estimate.html
17. Hanson, R.: The Age of Em: Work, Love, and Life when Robots Rule the Earth, 2nd ed (paperback) edn. Oxford University Press (2018)
18. Mercier, H., Sperber, D.: Why do humans reason? Arguments for an argumentative theory. Behavioral and Brain Sciences **34**(2) (2011). DOI 10.1017/S0140525X10000968
19. Nisbett, R.E., Wilson, T.D.: Telling more than we can know: verbal reports on mental processes. Psychological review **84**(3), 231 (1977)
20. Stumpf, E.: Förderung bei Hochbegabung. Kohlhammer Verlag (2011)
21. Naam, R.: Why AIs Won't Ascend in the Blink of an Eye - Some Math (2014). URL http://www.antipope.org/charlie/blog-static/2014/02/why-ais-wont-ascend-in-blink-of-an-eye.html
22. Lipp, M., Schwarz, M., Gruss, D., Prescher, T., Haas, W., Fogh, A., Horn, J., Mangard, S., Kocher, P., Genkin, D., et al.: Meltdown: Reading kernel memory from user space. In: 27th {USENIX} Security Symposium ({USENIX} Security 18), pp. 973–990 (2018)
23. Naam, R.: The singularity is further than it appears. Charlie's Diary (2014). URL http://www.antipope.org/charlie/blog-static/2014/02/the-singularity-is-further-tha.html
24. Ko, K.H.: Hominin interbreeding and the evolution of human variation. Journal of Biological Research-Thessaloniki **23**(1), 17 (2016)
25. Darwin, C.: The origin of species. PF Collier & son New York (1909)
26. Schrodinger, R., Schrödinger, E., Dinger, E.S.: What is life?: With mind and matter and autobiographical sketches. Cambridge university press (1992)
27. Sterling, B.: Schismatrix plus. Ace Books / Berkeley Publishing Group (1996)
28. Parker, M.P.: Intelligence and dysgenic fertility: Re-specification and reanalysis. Chrestomathy: Annual Review of Undergraduate Research at the College of Charleston **3**, 167–181 (2004)
29. Grace, K.: Algorithmic progress in six domains. Tech. rep., Machine Intelligence Research Institute (2013)
30. Tuomi, I.: The lives and death of Moore's Law. First Monday (2002)
31. Wikipedia contributors: Transistor Count. https://en.wikipedia.org/wiki/Transistor_count (2020). Downloaded on 24th Nov. 2020
32. Wikipedia contributors: Montecito (Processor). https://en.wikipedia.org/wiki/Montecito_(processor) (2020). Downloaded on 4th Dec. 2020
33. Waldrop, B.M.: The chips are down for Moore's law. Nature **530**, 145 (2016)
34. Kumar, S.: Fundamental limits to Moore's law. arXiv preprint arXiv:1511.05956 (2015)

35. Meieran, E.S.: 21st century semiconductor manufacturing capabilities. Intel Technology Journal **1**(8) (1998)
36. Steane, A.: Quantum computing. Reports on Progress in Physics **61**(2), 117 (1998)
37. Lalieu, M.L., Lavrijsen, R., Koopmans, B.: Integrating all-optical switching with spintronics. Nature communications **10**(1), 1–6 (2019)
38. Lin, Y.M., Dimitrakopoulos, C., Jenkins, K.A., Farmer, D.B., Chiu, H.Y., Grill, A., Avouris, P.: 100-ghz transistors from wafer-scale epitaxial graphene. Science **327**(5966), 662–662 (2010)
39. Landauer, R.: Irreversibility and heat generation in the computing process. IBM journal of research and development **5**(3), 183–191 (1961)
40. Rid, T.: Rise of the Machines: the lost history of cybernetics. Scribe Publications (2016)
41. McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E.: A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955 (Reprint). AI magazine **27**(4), 12–12 (2006)
42. McCulloch, W.S., Pitts, W.: A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics **5**(4), 115–133 (1943)
43. Turing, A.M.: I.—COMPUTING MACHINERY AND INTELLIGENCE. Mind **LIX**(236), 433–460 (1950). DOI 10.1093/mind/LIX.236.433. URL https://doi.org/10.1093/mind/LIX.236.433
44. Bradeško, L., Mladenić, D.: A survey of chatbot systems through a loebner prize competition. In: Proceedings of Slovenian Language Technologies Society Eighth Conference of Language Technologies, pp. 34–37 (2012)
45. Samuel, A.L.: Some studies in machine learning using the game of checkers. IBM Journal of research and development **3**(3), 210–229 (1959)
46. Newell, A., Shaw, J.C., Simon, H.A.: Report on a general problem solving program. In: IFIP congress, vol. 256, p. 64. Pittsburgh, PA (1959)
47. Winograd, T.: Procedures as a representation for data in a computer program for understanding natural language. Tech. rep., MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC (1971)
48. Tesauro, G.: Temporal difference learning and TD-Gammon. Communications of the ACM **38**(3), 58–68 (1995)
49. Pollack, J.B., Blair, A.D.: Why did TD-gammon work? In: Advances in Neural Information Processing Systems, pp. 10–16 (1997)
50. Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L., Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of go with deep neural networks and tree search. Nature **529**, 484–489 (2016). DOI 10.1038/nature16961
51. Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., Gale, J., Halpenny, M., Hoffmann, G., Lau, K., Oakley, C., Palatucci, M., Pratt, V., Stang, P., Strohband, S., Dupont, C., Jendrossek, L.E., Koelen, C., Markey, C., Rummel, C., van Niekerk, J., Jensen, E., Alessandrini, P., Bradski, G., Davies, B., Ettinger, S., Kaehler, A., Nefian, A., Mahoney, P.: Winning the DARPA Grand Challenge. Journal of Field Robotics (2006). Accepted for publication
52. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR 2011, pp. 1297–1304. Ieee (2011)
53. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., et al.: Building Watson: An overview of the DeepQA project. AI magazine **31**(3), 59–79 (2010)

54. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al.: Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal processing magazine **29**(6), 82–97 (2012)
55. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al.: Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv preprint arXiv:1712.01815 (2017)
56. Heule, M.J.: Schur number five. arXiv preprint arXiv:1711.08076 (2017)
57. Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W., Dudzik, A., Huang, A., Georgiev, P., Powell, R., Ewalds, T., Horgan, D., Kroiss, M., Danihelka, I., Agapiou, J., Oh, J., Dalibard, V., Choi, D., Sifre, L., Sulsky, Y., Vezhnevets, S., Molloy, J., Cai, T., Budden, D., Paine, T., Gulcehre, C., Wang, Z., Pfaff, T., Pohlen, T., Yogatama, D., Cohen, J., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Apps, C., Kavukcuoglu, K., Hassabis, D., Silver, D.: AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/ (2019)
58. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., Ronneberger, O., Bates, R., Žídek, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Potapenko, A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Steinegger, M., Pacholska, M., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D.: High Accuracy Protein Structure Prediction Using Deep Learning. https://predictioncenter.org/casp14/doc/CASP14_Abstracts.pdf (2020)
59. Rosenblatt, F.: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review **65**(6), 386 (1958)
60. Minsky, M., Papert, S.: An introduction to computational geometry. Cambridge tiass., HIT (1969)
61. Werbos, P.J.: Applications of advances in nonlinear sensitivity analysis. In: System modeling and optimization, pp. 762–770. Springer (1982)
62. Kelley, H.J.: Gradient theory of optimal flight paths. Ars Journal **30**(10), 947–954 (1960)
63. Platt, J.: Fast training of support vector machines using sequential minimal optimization. Advances in Kernel Methods—Support Vector Learning (pp. 185–208). AJ, MIT Press, Cambridge, MA (1999)
64. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. nature **323**(6088), 533–536 (1986)
65. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L.: Handwritten digit recognition with a back-propagation network. Advances in neural information processing systems **2**, 396–404 (1989)
66. Seewald, AK.: On the Brittleness of Handwritten Digit Recognition Models. ISRN Machine Vision **2012** (2012). DOI https://doi.org/10.5402/2012/834127
67. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
68. Shamsabadi, A.S., Sanchez-Matilla, R., Cavallaro, A.: Colorfool: Semantic adversarial colorization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1151–1160 (2020)
69. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2574–2582 (2016)
70. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
71. Wikipedia contributors: Comparison of deep-learning software. https://en.wikipedia.org/wiki/Comparison_of_deep-learning_software (2020). Downloaded on 4th Dec. 2020

72. Sharma, N., Smeets, B., Tryggestad, C.: The decoupling of GDP and energy growth: A CEO guide. The McKinsey Quarterly (2019)
73. Gordon, R.J.: The rise and fall of American growth: The US standard of living since the civil war, vol. 70. Princeton University Press (2017)
74. Crivellato, E., Ribatti, D.: Soul, mind, brain: Greek philosophy and the birth of neuroscience. Brain Research Bulletin **71**(4), 327 – 336 (2007). DOI https://doi.org/10.1016/j.brainresbull.2006.09.020. URL http://www.sciencedirect.com/science/article/pii/S036192300600298X
75. Huxley, T.H.: On the hypothesis that animals are automata, and its history. Collected essays **1** (1874)
76. Wikipedia contributors: Maillardet's automaton. https://en.wikipedia.org/wiki/Maillardet%27s_automaton (2020). Downloaded on 4th Dec. 2020
77. Wikipedia contributors: Digesting Duck. https://en.wikipedia.org/wiki/Digesting_Duck (2020). Downloaded on 4th Dec. 2020
78. Wikipedia contributors: Euphonia (Sprachmaschine). https://de.wikipedia.org/wiki/Euphonia_(Sprachmaschine) (2020). Downloaded on 4th Dec. 2020
79. Penrose, R.: The emperor's new mind. RSA Journal **139**(5420), 506–514 (1991)
80. Adams, B., Petruccione, F.: Quantum effects in the brain: A review. AVS Quantum Science **2**(2), 022,901 (2020)
81. Kandel, E.R., Schwartz, J.H., Jessell, T.M., of Biochemistry, D., Jessell, M.B.T., Siegelbaum, S., Hudspeth, A.: Principles of neural science, 4th Edition. McGraw-Hill New York (2000)
82. Heindl, A., Schepelmann, M., Nica, R., Ecker, R., Pietschmann, P., Seewald, A.K., Thalhammer, T., Ellinger, I.: Toward the automated detection and characterization of osteoclasts in microscopic images. In: Principles of Osteoimmunology, pp. 31–57. Springer (2016)
83. Seewald, A.K.: Revisiting End-to-end Deep Learning for Obstacle Avoidance: Replication and Open Issues. In: ICAART (2), pp. 652–659 (2020)
84. Seewald, A.K., Cypser, J., Mendenhall, A., Johnson, T.: Quantifying phenotypic variation in isogenic Caenorhabditis elegans expressing Phsp-16.2:: gfp by clustering 2D expression patterns. PloS one **5**(7), e11,426 (2010)
85. Hofstadter, D.R.: I am a strange loop. Basic books (2007)
86. Sawyer, R.J.: WWW: wake. Penguin (2009)
87. Hogan, J.P.: The Two Faces of Tomorrow. Baen Books (1997)
88. Hogan, J.P.: Mind Matters Exploring the World of Artificial Intelligence. Del Ray / Ballantine Publishing Group (1998)
89. Hameroff, S., Penrose, R.: Reply to seven commentaries on "Consciousness in the universe: Review of the 'Orch OR' theory". Physics of Life Reviews **11**(1), 94–100 (2014). DOI 10.1016/j.plrev.2013.11.013