13. The encroachment of artificial intelligence: Timing and prospects

Richard Granger, Dartmouth College

This chapter describes several instances of artificial intelligence (AI), artificial neural network and machine learning systems that are judged to be highly successful. It also highlights the shortcomings of these systems to explain their limitations. Through examples such as so-called self-driving cars, image recognition, handwriting analysis and digital virtual assistants like Siri, the chapter explores the ways in which AI is both like and unlike human intelligence. It clarifies ways in which AI will and will not be useful in various workplaces. It also examines capabilities of humans that are likely to outpace AI for some time, and therefore may remain critical factors in employment practice.

Introduction

Artificial intelligence (AI) systems are steadily growing as tools to accomplish specified tasks, increasingly at the expense of jobs that would otherwise be carried out by human workers. How can people stay relevant and employed?

This chapter focuses on shortfalls of AI, explaining limitations likely to persist for many years. It clarifies ways in which AI will and will not be useful in various workplaces. It also examines capabilities of humans that are likely to outpace AI for some time, and therefore may remain critical factors in employment practice.

To that end, it focuses on ways in which AI – and related advances in machine learning and artificial neural networks (ANN) – is like, and quite unlike, human intelligence. How are the architectures – the innards – of ANNs like, and unlike, actual networks of neurons, i.e. brain circuits? What behavioural or computational abilities arise from ANNs, and how are they like and unlike human behavioural and computational abilities?

Similar and different underlying architectures

Artificial and natural neural networks

How are artificial "neural networks" similar to, and different from, actual brain circuitry? AI systems have shown an impressive ability to play difficult games, perform language translation, find patterns in complex data and many other seemingly human-level behaviours. These systems are often designated as "brain-like", or "brain-inspired", i.e. based on or derived from design principles of human brains.

Given the impressive accomplishments of these systems, it is perhaps understandable, even flattering, to connect them to human brain designs. Where do such claims come from? What characteristics of brains do these artificial systems actually exhibit?

Brains differ from standard (non-AI) computers in myriad ways. Among other things, brains learn and retain information over decades, whereas computers do not. Further, computers are predominantly serial whereas brains exhibit large-scale parallelism.

Surprisingly few characteristics of brains that might have become standard ANN properties have been adopted. These include long-term learning and parallelism, as well as the neural property of only producing simple computations such as addition and multiplication. In this way, the engines of the brain (neurons) contrast with computers that have extensive mathematical abilities.

Today, ANNs comprise much (though not all) of what is often more broadly termed machine learning and AI. They are composed of simple computing units, acting in parallel and learning from data. This indeed makes ANNs far more "brain-like" than standard computers.

However, these points of similarity between artificial networks versus actual brains represent a surprisingly small fraction of brain characteristics. These neural and brain-circuit properties (e.g. location and temporal differences between excitatory and inhibitory cells) can seem insignificant. Yet the capabilities of ANNs versus real brains are vastly different. It is not yet known what cognitive abilities arise from what features of our brains. Many AI and ANN researchers have shown that incorporation of additional actual brain properties may substantially, even drastically, alter and enhance the capabilities of ANNs.

Many AI researchers are studying how to overcome the shortfalls of AI systems compared to human abilities. Although AIs outperform many human abilities, these strongly centre on circumscribed tasks such as game-playing and online product recommendations. Interestingly, these closely correspond to the type of task for which computers have always outperformed humans: large-scale numeric calculation and data analysis.

AIs excel at tasks that can be reduced to pre-specified outcomes such as either winning or losing a game of chess or Go. It remains unclear whether more open-ended intelligent behaviours will be susceptible to similar approaches. Correspondingly, AIs still are markedly inferior to humans in reasoning, lifelong learning, common sense and much more.

Brain algorithms are intrinsically parallel; other algorithms typically are not

Unlike many algorithms, brain algorithms are intrinsically parallel. Computational steps can only be carried out in parallel if later steps in a process do not depend on previous steps. This avoidance of “serial dependency” is at the core of parallelism.

In a large computational set of operations, any given step in the process (say, step 146) may depend on prior steps (e.g. steps 26, 91 and 108) to have produced their partial outputs. The essence of “serial” computation is the dependency of later operations on earlier operations. All standard computers are intrinsically serial.

Can these serial dependencies somehow be turned into independent parallel operations? This is a present-day, much-researched problem that remains unsolved. Despite many specialised approaches, there is no general method for taking the dependencies in serial operations and somehow “parallelising” them.

A first step in understanding brains is to understand their already-parallel methods – their parallel algorithms – rather than assuming that computational serial methods can be rendered parallel.

Many hypothetical models of brain systems have been characterised in terms of their parallel and serial components. A prominent example includes “unsupervised” categorisation methods, which are predominantly intrinsically parallel – not “parallelised” from serial steps; these have been proposed as partial models of some cortical operations. Another such example is “reinforcement learning” (RL) systems, which include highly parallel operations. These have been widely proposed as partial models of the basal ganglia structures in brains.

Some additional related computational operations are also intrinsically parallel. For instance, most of what a search engine does is parallel: I type in “antiviral” to Google. It can parcel out the task to millions of independent computers to search through separate, independent, parts of the Internet.

After all those independent jobs return their findings, a non-parallel job must put them all together and then order them from first to last. However, most of the task is intrinsically parallel: each separate location in the Internet can be looked at separately during the search. Since none of them depends on others, all these separate searches can be done simultaneously and independently with enough separate computers to assign in parallel to this task.

Parallel computers are becoming increasingly important and prevalent. Such computer hardware (such as supercomputers, clusters, GPUs) contain substantial numbers of computing units (think neurons). These can carry out their operations independently of each other, and therefore can operate simultaneously. Many parallel methods (such as search) are typically implemented onto parallel hardware to speed up their operation.

Remarkably, most ANN operations, especially the still-prevalent “supervised” systems such as “backpropagation” (and “multi-layer perceptrons”, more generally), are not entirely parallel. Consequently, they still require expensive hardware to run.

Interestingly, vast swaths of land around major hydroelectric dam sites are owned by the few major tech firms. Most of these firms specialise in AI, including Google (Alphabet), Apple, Amazon, Facebook, Netflix, Microsoft and IBM.

Why this strange connection between computer tech companies and hydroelectric dams? Their "server farms", i.e. stations of large numbers of computer clusters, run so hot that their power and cooling costs require plentiful power sources. In other words, AI is not cheap or prevalent as yet: much of it still requires extreme resource usage just to run at all (Jones, 2018[1]; Pearce, 2018[2]).

Similar and different behaviours and "cognitive" capabilities

How is someone's intelligence or aptitude assessed for a task? Typically, specific tests designed to identify traits and abilities have been found to correlate with humans who previously performed well in the task or job of interest.

Tests of this kind omit enormous sets of actual worker characteristics desired by an employer. This is because the worker is assumed to be human with standard human capabilities. These capabilities are taken for granted, and not explicitly tested.

In other words, humans are ostensibly tested for their necessary skills. However, many crucial skills are simply assumed since all humans possess them. One such set of skills is the ability to not commit catastrophic and unexpected errors. The four examples below – i) self-driving cars, ii) image recognition, iii) handwriting analysis and iv) digital virtual assistants – explore these concerns.

"Self-driving" cars

The term "self-driving" car is a misnomer. The numeric multi-tier "levels" scale, outlined below, roughly describes increasingly difficult abilities rather than levels of autonomy (NHTSA, 2018[3]; SAE, 2018[4]):

0. No automation: i.e. normal cars; all driving responsibilities resting with human drivers.
1. Driver assistance: human drivers are assumed to be responsible for all tasks, and may be assisted by a system that partially performs steering, *or* acceleration/deceleration.
2. Partial automation: similar to "1", humans are assumed to be responsible for all tasks, and may be assisted by a system that partially performs steering *and* acceleration/deceleration.
3. Conditional automation: the artificial system takes over most aspects of driving, including steering and acceleration/deceleration. Yet the human driver is expected to "respond appropriately" to a request to take over. In other words, ongoing vigilance is still required on the part of the human driver.
4. High automation: the artificial system takes over most aspects of driving. If the human driver does not respond appropriately to a request to take over, the system can pull over to stop the vehicle.
5. Full automation: the artificial system performs any and all driving tasks of a human driver, with no intervention required from the human.

No current products go beyond "Level 3". In other words, they all depend on the human driver to take over whenever called to do so by the AI system. Higher-level products have often been predicted, but many experts explain they remain a long way off.

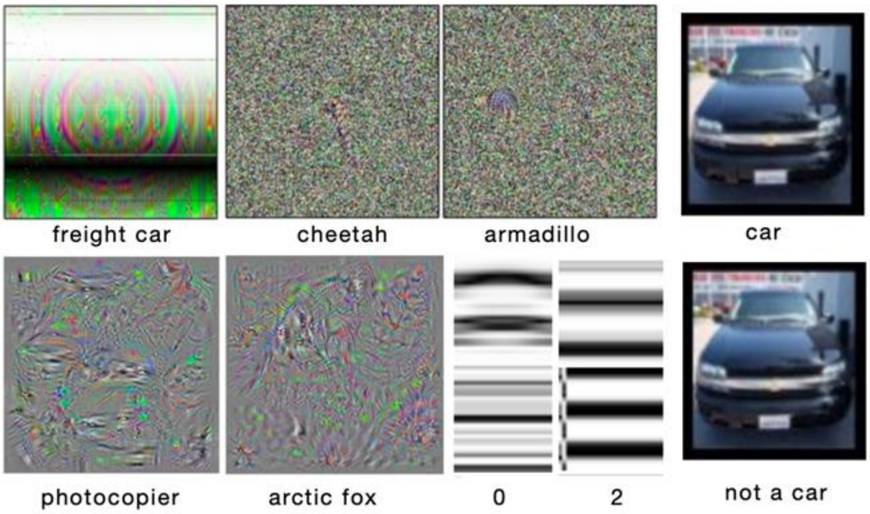
When "Level 3" cars have accidents, the accidents themselves are not the primary issue. (Humans, too, have accidents.) Rather, it is the novelty of Level 3 accidents that is striking. The accidents involve things that, again, no one even thought to test for because they are so far outside the realm of human experience. In one case, for instance, a Level 3 car drove at high speed into the side of a truck. To the AI system, the truck somehow looked like the sky (Tian et al., 2018[5]; Boudette, 2019[6]; TKS Lab, 2019[7]).

There are multitudes of such potential unexpected and catastrophic errors lurking in AIs. These errors are not tested for in advance, and current tests consistently fail to anticipate them. Humans would not, for instance, accidentally insert an iPhone into a coffee machine or ignite their office chair.

Image recognition

Equally compelling examples of catastrophic errors occur in image recognition. The consequences are typically (though not always) less dramatic than Level 3 car accidents. However, the errors are every bit as illuminating. Other things besides images that fool humans can fool AI image experts. Figure 13.1 shows a set of "adversarial" images (Szegedy et al., 2014[8]; Kurakin, Goodfellow and Bengio, 2016[9]) that are judged, by these expert AI systems, to be the object labelled below them. For example, one pixilated multicolour image was identified by the AI as a cheetah. Another such pixilated image was identified as an armadillo. To humans, these images do not remotely resemble animals or objects.

Figure 13.1. Adversarial images showcase judgement errors of AI systems



Source: (Nguyen, Yosinski and Clune, 2015[10]).

The designers of these systems did not think to test for such errors, which illustrates the range of the problem. It appears not to be practicable to anticipate all such errors; the range of misreads is so broad and unexpected that it cannot readily be predicted.

Why do these strangely inhuman errors arise at all? If AI systems performed somewhat like humans do, they might fall short in certain ways, much as some humans exhibit capabilities that other humans may lack. However, AI systems are not carrying out the kinds of operations that humans do.

Although AI systems are touted as "human-like" and "brain-like", there is overwhelming evidence by AI experts indicating this is not so. For example, the widely used "supervised" learning systems of most ANNs have sometimes been suggested as relating to operations of the human neocortex. However, brain projections do not contain "error correction" signals of the kind required by supervised systems.[1]

A retrospective case: Handwriting recognition

Building on two examples of the non-human nature of AIs (operating cars and recognising images), this section considers the "retrospective" example of handwriting recognition. It illustrates the perspective of time. When errors occur, how long does it take the field to recover from them?

Early focus on handwriting in the 1990s

In the 1990s, handwritten text recognition presented a major challenge to AI. Major conferences were held, focusing solely on handwriting recognition. In the United States, entire funding programmes of the Defense Advanced Research Projects Agency supported the work, and entire divisions of the National Institute of Standards and Technology developed datasets for it.

By the late 1990s, the field believed it was succeeding (LeCun, Bottou and Haffner, 1998[11]; Von Ahn et al., 2003[12]). Publications routinely highlighted their "correct hit rate" achievements, with suspiciously precise-seeming values such as 91.4% correct. These assessments assumed that the phrase "handwriting recognition" was self-explanatory.

Development of CATCHAs in the 2000s

However, in the early 2000s, new data appeared in the form of "Completely Automated Public Turing test to tell Computers and Humans Apart" (CAPTCHAs) (Von Ahn et al., 2003[12]). Websites began using these forms of distorted letters (now ubiquitous) to determine whether a purported user was a human or an AI. Humans could see the letters and numbers in CAPTCHAs effortlessly.

However, handwriting-recognition AI systems performed abysmally on CAPTCHAs. A few papers described promising recognition successes on limited databases but failed to generalise to other datasets. The supposedly highly successful field of handwriting recognition went almost entirely dry for more than ten years. Dozens of claims were made that CAPTCHAs had been cracked; these were repeatedly refuted.

Finally, 14 years later, a publication showed an approach that could reliably address many popular CAPTCHA systems (Bursztein et al., 2014[13]). It took additional years for other such reports to emerge; some were still considered to be sufficiently noteworthy that they appeared in prestigious journals (George et al., 2017[14]).

In sum, a problem introduced in 2000, into a supposedly successful field, essentially crushed that field until the mid to late 2010s. All of this took multiple lifetimes in terms of typical technology cycles.

The need for broader assessment of systems

Today, the stakes are similar. The predominant focus of researchers on the statistics of huge data, such as games and shopping recommendations, artfully alters the metric for success. They are not addressing open-ended problems, i.e. the problems that humans typically face. Rather, these approaches aim at achieving known metrics such as game wins.

These closed-ended tasks are being won by massive memorisation and processing of millions of instances (Serre, 2019[15]). The claims of game-playing systems, for instance, refer to software trained on the equivalent of the entire lifetimes of imagined hundreds of thousands of human players. By contrast, humans perform highly complex tasks of recognition, retrieval, decision and inference, after learning on comparatively miniscule quantities of data, many orders of magnitude less than the artificial systems require.

Without the specifications of actual human behaviour, it can be all too easy to imagine researchers are formally addressing a task such as handwriting recognition, or chess or Go.

The lesson is not being learnt. Tasks are carefully steered away from far-reaching human abilities, focusing instead on data memorisation mixed with slight generalisation. By this nostrum, a stream of attractive successes are toted up, largely disregarding failures and shortfalls (Serre, 2019[15]).

Many in the field are mindful of the situation. Researchers are striving for applicable metrics that could assess systems more broadly, seeking "feasible and reasonable" tests to which a system could be subjected. A range of views on this approach is worth further pursuit (Legg and Hutter, 2007[16]; Dowe and Hernandez-Orallo, 2012[17]; Hernandez-Orallo, Dowe and Hernandez-Lloreda, 2014[18]).

“Digital virtual assistants”

This section turns to a set of examples perhaps somewhat closer to direct experience: the performance of Siri (and corresponding conversational AIs, such as Alexa, etc.).

Siri and its cousins, now many years old, are still revolutionary and impressive accomplishments. Their recognition of spoken English and several other languages is still among the more skilful achievements of commercial AIs. Yet Siri makes jarring errors that even a young child would not make. Why is this so? Why do Siri’s errors not match those that people might produce?

Indeed, how does one know when Siri has made an error at all? One way to address this is to envision a rigorous system for recognising and cataloguing such errors. After all, if humans cannot catch the errors, how could they possibly design systems that produce fewer of them?

Importantly, it is not possible, even in principle, to construct any rigorous system for recognising Siri’s errors. This is because the measure of “when something is an error” is solely empirical: a human must judge that the response (somehow) does not make sense. By Siri’s internal logic, of course, the response was somehow the correct output computed from the input she received.

Empirically, then, a human evaluator is required to judge whether or when Siri errs. This is radically different from the corresponding circumstances of other systems used in offices. If a corporation purchases a photocopying machine or an assembly line system, they come with specifications: careful, relatively precise descriptions of what the system will do. Deviations from these specs are errors. Even software, including complex software, comes with specifications. There are explicit instructions for its use, and careful characterisations of its corresponding prescribed behaviours.

Siri comes with no such specifications. In fact, its creators cannot, with any precision at all, produce such specifications. Errors, then, are not deviations from (non-existent) specifications. What counts as an error? Only what humans notice, after the fact.

This is the crucial difference between all AI systems and their hybrids, compared to all other contemplated office or workplace systems. With few or no specifications, errors cannot be reliably predicted in advance (Granger, 2020[19]).

Conclusions

AI system problems are real, and sometimes nefarious

It is a time of substantial upheaval in the fields of AI, machine learning and neural networks. The normally well-regarded journal *Nature* recently published a report co-authored by researchers at Google, describing testing of an AI system for medical screening of breast cancer mammograms (McKinney et al., 2020[20]). It then published a critique by a group of AI experts, who argued “the absence of sufficiently documented methods and computer code underlying the study effectively undermines its scientific value” (Haibe-Kains et al., 2020[21]).

The original authors then issued a brief counter-reply, asserting they would not release all of the code or data used to obtain the results. They argued that “much of the remaining code … [is] of scant scientific value and limited utility to researchers outside our organisation.” They further stated that releasing the system could risk it being treated as “medical device software” and “could lead to its misuse” (McKinney et al., 2020[22]).

These claims seem transparently spurious for two reasons. First, the requested code is indeed of direct interest to the researchers attempting to evaluate and replicate the medical claims. Second, the code could be issued solely to AI researchers and clearly labelled as experimental and expressly not for medical use.

Corporations are similarly reluctant to release data around the pursuit of autonomous driving, medical diagnostics, and many other systems with potentially widespread and dangerous impacts. This continues to make it impossible in most cases for researchers to evaluate the claims of such systems. When the methods are hidden, they cannot in any way be seriously evaluated. In such cases, they simply remain unsubstantiated claims, and should be treated as such.

Even supposedly non-profit research organisations have refused to make their research code and materials available for evaluation by scientific researchers. The perhaps inaptly named "Open AI" switched from being a non-profit to a "capped profit". This turns out to mean a for-profit corporation that says it sets a 100x limit on investors' returns. For comparison, early investors in Google have received roughly a 20x return.

OpenAI publicised GPT-2 and later GPT-3 that can write impressive-seeming text stories. These initially were presented as literally "understanding human language". However, the code producing these purported wonders was secret. OpenAI, despite its name, would not release the code (Gershgorn, 2020[23]).

Even well-respected academic researchers could not access the code to test the company's claims. Eventually, some researchers did (not via OpenAI itself), and published findings that drastically call into question the code's "understanding" of language. A sample failed story, for example, left most readers befuddled at what was being said[2] (Marcus and Davis, 2020[24]; Marcus and Davis, 2020[25]). Extended discussions of other failures of common sense in present-day AI are well presented in Marcus and Davis (2019[26]).

A statement may be syntactically correct and seemingly coherent, and yet not make any sense. Yet, like Siri, this is not an example of a system failing in the typical sense. These systems have opaque designs, such that their failures cannot be rigorously predicted. Even after the fact, the failures cannot be cogently systematised.

How can we be more informed consumers and testers of proposed AI systems?

What can be done to ensure that humans continue to participate and thrive in work environments targeted by AI systems? Despite enormous publicity and fanfare, AIs do not appear ready to take over jobs that require human language usage and common sense. AIs applying for such jobs will continue to be blindsided by many real-world situations, like trucks on the highway. They are even more vulnerable to directed attacks such as adversarial inputs.

Recognise shortcomings of AI systems

It remains unclear whether specific examples of the kind described here could be generalised in a way to thoroughly identify likely AI errors. A systematic effort to collect known errors can assist in individual assessments of proposed AI systems (by checking for known errors). However, it is not known whether these will lead to systems that are so improved that such errors no longer occur.

Some companies are working to develop products aimed directly at overcoming these AI shortcomings. A "hybrid" approach, for example, incorporates higher-level symbol-manipulation operations with lower level ANN systems for statistical handling of big data (Marcus and Davis, 2019[26]). This approach includes symbol-manipulation systems, such as "expert systems" from past successful AI efforts. As such, they allow specific "common sense" rules, such as "if this, then that", which systems can use to evaluate possible actions.

Hybrid systems may not be the answer, but new approaches of some kind must be brought to bear. At the very least, AI “experts” could increasingly be trained to assess proposed systems for known errors. They could also acknowledge the lack of generalisation that is a reliable hallmark of extant systems.

Train AI experts and customers to look for known errors

Customers of hybrid systems could perhaps usefully be trained to query experts with these examples of widespread AI shortcomings in mind. If potential AI customers are presented with sample errors such as these, they may become increasingly well-informed users. As a result, they may become qualified to ask AI providers and experts about the precise limits of a candidate system under specific conditions. With the intended tasks of the system specified, experts can then be asked to consider an expanded set of tasks. The aim would be to clarify which generalisations of the AI system can and cannot be counted on.

These conclusions indicate how difficult it may be to define any remotely “complete” set of tasks that a buyer, or an AI expert, could use to identify all potentially relevant AI errors. It is far from obvious how to pin down even a small representative set of such tests.

The four categories of examples described in this manuscript suggest initial starting points for these common areas of supposed expertise (self-driving cars; image recognition; handwriting; digital virtual assistants). These and many other available instances of AI hard limitations should be made available to entities looking to acquire and use AI systems.

It will likely be beneficial for the pre-planned domain of the intended AI system to draw on these starting points to formulate examples. For instance, in each example, a few themes appear to emerge. Limitations to AI systems are described in terms that appear technical and detailed but in practice are vague with respect to what a user can expect (as in self-driving cars, handwriting, image recognition). Operations not explicitly tested are not highlighted for their possible divergence from reasonable behaviour (all categories display this shortcoming).

Use concrete vignettes to avoid incorrect inferences

In addition, AI systems are presented such that an intended user may infer abilities that a human would exhibit in a job, but that the AI system is not at all guaranteed to do. For example, the differences between “Level 2” and “Level 3” in self-driving cars are described by long, and highly detailed and technical descriptions. However, these can readily mislead readers into making inferences not promised in the specification.

The leap between Level 2 and Level 3 is supposedly separated by a system in which “the human monitors the driving environment” (Level 2) vs. “the automated system monitors the driving environment” (Level 3). Yet, despite the addition of multiple sensors in the latter system, the human is nonetheless still required to continuously monitor all conditions, and be prepared to take control at a moment’s notice.

How might users detect such potential misreads in advance? One approach is to take the specification of the task, and identify specific, concrete vignettes in which a difficulty might arise. These difficulties could be either for the human in a job, or for the AI system supposedly performing that job. Ask specific questions about what the human might be expected to do, if, say, a shipment does not arrive on schedule. AI system descriptions do not typically volunteer such examples.

In each case, the more scenarios that are specifically inquired about, the more likely that system errors may be identified. This approach may appear too cumbersome or piecemeal, but it is exactly this characteristic that may make it useful.

The approach may seem piecemeal because it appears to lack an undergirding principle connecting different proposed vignettes. Indeed, AI systems do not yet have anything resembling complete principles for behaviours in complex settings. This is why a car may crash at full speed into a wall, or an image that looks like white noise may be labelled as an armadillo, with high (but erroneous) confidence.

A human would handle edge effects with “common sense”, but the AI system may prove erratic. Indeed, it may be erratic in any untested situation. Testers would do well to think of tasks where a human might wince or express concern, but finally react sensibly. In these situations, AI may well fail the test.

AI systems yet to be

An oft-told adage is that present-day AI approaches are like climbing a tree, or even flying a helicopter, with the aim of reaching the moon. They can cite measurable ongoing progress: they do indeed keep climbing higher, closer to their goal. Yet it will require utterly distinct understandings and methods to go beyond the current achievements.

In the long run, there is no principled reason why artificial systems cannot duplicate, and exceed, human abilities. Indeed, current human abilities already arise from physical machines – our brains; they simply are made of meat rather than metal. In describing the original purpose of AI, John McCarthy said, “Every aspect of learning or any other feature of intelligence can be so precisely described that a machine can be made to simulate it” [McCarthy et al., 1955 in (McCarthy et al., 2006[27])].

The field of AI is not currently “precisely describing” the features of intelligence. Quite the opposite: such features are still highly elusive. This definition of AI does not emphasise analysing large datasets or commercial products per se. Rather, it focuses on the scientific objective of understanding the mind, and the brain that produces the mind. This may appear to be a too-theoretical pursuit, but it is the most pragmatic path to follow: to outperform humans, one must first equal them.

AIs will eventually outperform humans but are nowhere close to doing so. AIs currently outperform humans in much the way that computers have always outperformed humans: in large-scale numeric calculation and data analysis. Correspondingly, AIs still wildly underperform humans in reasoning, lifelong learning, common sense and much more.

As the “precise descriptions” of these human abilities are achieved, the inception of true intelligences will also come closer. Crucial information will come from neuroscience, psychology, computer science, mathematics and other related fields. Just as flying machines were based on principles of aerodynamics used by flying animals, intelligent machines will arise from understanding the principles of intelligence. Great advances have been made towards this aim, and more will come. In the meanwhile, human workers significantly outpace AIs in their judgements and practicality.

When AIs do come to verge on human common sense, the questions of their industrial utility will be even more urgent. Jobholders have been repeatedly infringed on in the past, in times of economic and societal reorganisation. All such previous upheavals have entailed humans taking the jobs of other humans, but the resulting instabilities were nonetheless real. As AIs advance beyond their current limitations, continued threats to stable human employment will recur. Much work is yet to be done to address the future livelihoods of humans in an increasingly artificially intelligent world.

References

Boudette, N. (2019), “Despite high hopes, self-driving cars are ‘Way in the future’”, 17 July, New York Times. [6]

Bursztein, E. et al. (2014), “The end is nigh: Generic solving of text-based CAPTCHAs”, presented at WOOT 14, San Diego, CA, 19 August, https://www.usenix.org/conference/woot14/workshop-program/presentation/bursztein. [13]

Chandrashekar, A. and R. Granger (2012), “Derivation of a novel efficient supervised learning algorithm from cortical-subcortical loops”, *Frontiers in Computational Neuroscence*, Vol. 5, http://dx.doi.org/10.3389/fncom.2011.00050. [31]

Dowe, D. and J. Hernandez-Orallo (2012), “IQ tests are not for machines, yet”, *Intelligence*, Vol. 40, pp. 77-81, https://doi.org/10.1016/j.intell.2011.12.001. [17]

George, D. et al. (2017), “A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs”, *Science*, http://dx.doi.org/10.1126/science.aag2612. [14]

Gershgorn, D. (2020), “GPT-3 is an amazing research tool. But OpenAI isn’t sharing the code”, *Medium,* 20 August, https://onezero.medium.com/gpt-3-is-an-amazing-research-tool-openai-isnt-sharing-the-code-d048ba39bbfd. [23]

Granger, R. (2020), “Toward the quantification of cognition”, *arXiv*, Vol. 2008.05580, https://arxiv.org/abs/2008.05580. [19]

Haibe-Kains, B. et al. (2020), “Transparency and reproducibility in artificial intelligence”, *Nature*, Vol. 546, pp. E-14–E-16, https://doi.org/10.1038/s41586-020-2766-y. [21]

Hernandez-Orallo, J., D. Dowe and M. Hernandez-Lloreda (2014), “Universal Psychometrics”, *Cognitive Systems Research*, Vol. 27, pp. 50-74, https://doi.org/10.1016/j.cogsys.2013.06.001. [18]

Jones, N. (2018), “How to stop data centres from gobbling up the world’s electricity”, *Nature*, Vol. 561, pp. 163-166, https://doi.org/10.1038/d41586-018-06610-y. [1]

Kurakin, A., I. Goodfellow and S. Bengio (2016), “Adversarial examples in the physical world”, *arXiv*, Vol. 1607.02533, https://arxiv.org/abs/1607.02533. [9]

LeCun, Y., L. Bottou and P. Haffner (1998), “Gradient-based learning applied to document recognition”, *Proceedings of the IEEE*, Vol. 86/11, pp. 2278-2324, http://dx.doi.org/10.1109/5.726791. [11]

Legg, S. and M. Hutter (2007), “Universal intelligence: A definition of machine intelligence”, *Minds and Machines*, Vol. 17/4, pp. 391-444, http://dx.doi.org/arXiv:0712.3329. [16]

Lillicrap, T. et al. (2020), “Backpropagation and the brain”, *Nature Reviews Neuroscience*, Vol. 21, pp. 335-346, https://doi.org/10.1038/s41583-020-0277-3. [30]

Marblestone, A., G. Wayne and K. Kording (2016), “Toward an integration of deep learning and neuroscience”, *Frontiers in Computational Neuroscence*, Vol. 10/94, https://doi.org/10.3389/fncom.2016.00094. [29]

Marcus, G. and E. Davis (2020), “Experiments testing GPT-3’s ability at commonsense reasoning”, Department of Computer Science, New York University, https://cs.nyu.edu/faculty/davise/papers/GPT3CompleteTests.html. [25]

Marcus, G. and E. Davis (2020), “GPT-3, Bloviator: OpenAI’s language generator has no idea what it’s talking about”, *MIT Technology Review* 22 August, https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/. [24]

Marcus, G. and E. Davis (2019), *Rebooting AI*, Penguin Random House, New York. [26]

McCarthy, J. et al. (2006), “A proposal for the Dartmouth summer research project on artificial Intelligence, August 31, 1955”, *AI Magazine*, Vol. 27/4, p. 12, https://doi.org/10.1609/aimag.v27i4.1904. [27]

McKinney, S. et al. (2020), “Reply to: Transparency and reproducibility in artificial intelligence”, *Nature*, Vol. 586, pp. E-17–E-18, https://doi.org/10.1038/s41586-020-2767-x. [22]

McKinney, S. et al. (2020), “International evaluation of an AI system for breast cancer screening”, *Nature*, Vol. 577/7788, pp. 89-94, https://doi.org/10.1038/s41586-019-1799-6. [20]

Nguyen, A., J. Yosinski and J. Clune (2015), *Deep neural networks are easily fooled: High confidence predictions for unrecognizable images*. [10]

NHTSA (2018), “A framework for automated driving system testable cases and scenarios”*, Report*, No. DOT HS 812 623, National Highway Traffic Safety Administration, Washington, DC, https://www.nhtsa.gov/sites/nhtsa.dot.gov/files/documents/13882-automateddrivingsystems_092618_v1a_tag.pdf. [3]

Pearce, F. (2018), “Energy hogs: Can world’s huge data centers be made more efficient?”, *Yale Environment*, Vol. 350, https://e360.yale.edu/features/energy-hogs-can-huge-data-centers-be-made-more-efficient. [2]

Rodriguez, A., J. Whitson and R. Granger (2004), “Derivation and analysis of basic computational operations of thalamocortical circuits”, *Journal of Cognitive Neuroscience*, Vol. 16, pp. 856-877, http://dx.doi.org/10.1162/089892904970690. [28]

SAE (2018), “Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles”*, Revised report*, No. J3016_201806, SAE International, Warrendale, PA, https://www.sae.org/standards/content/j3016_201806/. [4]

Serre, T. (2019), “Deep learning: The good, the bad, and the ugly”, *Annual Review of Vision Science*, Vol. 5, pp. 399-426, https://doi.org/10.1146/annurev-vision-091718-014951. [15]

Szegedy, C. et al. (2014), “Intriguing properties of neural networks”, *arXiv*, Vol. 1312.6199, http://dx.doi.org/arxiv.org/abs/1312.6199. [8]

Tian, Y. et al. (2018), “DeepTest: Automated testing of deep-neural-network-driven autonomous cars”, *arXiv*, Vol. 1708.08559, http://dx.doi.org/arXiv:1708.08559. [5]

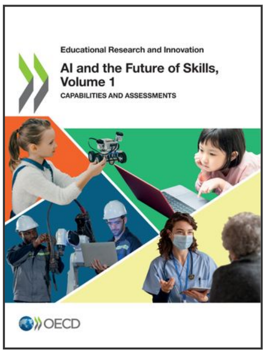
TKS Lab (2019), *Experimental Security Research of Tesla Autopilot*, Tencent Keen Security Lab, Shenzen, https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf. [7]

Von Ahn, L. et al. (2003), *CAPTCHA: Using hard AI problems for security*, Springer, http://dx.doi.org/10.1007/3-540-39200-9_18. [12]

Notes

[1] Several recent efforts have tried to demonstrate an abstract mathematical relationship between some operations that brains may actually do and the back-propagated error correction of ANNs. Such convoluted arguments may serve in part as reminders of the radical differences between ANNs and brains. They emphasise the comparative lack of research on how the desired computational ends may be instead achieved via different algorithms that do not rely on this form of error correction (Rodriguez, Whitson and Granger, 2004[28]; Chandrashekar and Granger, 2012[31]; Marblestone, Wayne and Kording, 2016[29]; Lillicrap et al., 2020[30]).

[2] One sample failed story produced the following text: “At the party, I poured myself a glass of lemonade, but it turned out to be too sour, so I added a little sugar. I didn’t see a spoon handy, so I stirred it with a cigarette. But that turned out to be a bad idea because it kept falling on the floor. That’s when he decided to start the Cremation Association of North America, which has become a major cremation provider with 145 locations.” This example is also discussed in Chapter 12.


Educational Research and Innovation
AI and the Future of Skills, Volume 1
CAPABILITIES AND ASSESSMENTS
OECD

From:

AI and the Future of Skills, Volume 1

Capabilities and Assessments

Access the complete publication at:
https://doi.org/10.1787/5ee71f34-en

Please cite this chapter as:

Granger, Richard (2021), “The encroachment of artificial intelligence: Timing and prospects”, in OECD, *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/e39255bb-en

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area. Extracts from publications may be subject to additional disclaimers, which are set out in the complete version of the publication, available at the link provided.