THE INDIA FORUM
A JOURNAL-MAGAZINE ON CONTEMPORARY ISSUES

August 10, 2021

Artificial Intelligence and its Discontents-I

By: Venu Madhav Govindu

Artificial Intelligence has exploded onto the world in recent years. Part-I of this review essay examines the nature and implications of contemporary AI. Next week, Part-II will address the multiple crises engendered by AI's rapid and widespread deployment.

In recent years, Artificial Intelligence or AI has leapfrogged from the research laboratory into many aspects of our lives. Face recognition, keyboard prompts, recommendations for what to buy on Amazon or whom to follow on Twitter, and text translation or searching for similar images are just some of the AI techniques that have become commonplace. At one remove from the ordinary user, AI is also being deployed in areas as diverse as radiological diagnostics, pharmaceutical drug design and drone navigation. It is little wonder then that AI is a new buzzword for our times and seen as a portal to a dramatic future.

Interest in automaton replicating human capabilities is age-old, but the discipline of AI can be dated with precision to a summer research project conceived by John McCarthy and others in 1956. These founding figures started work based on "the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it". The objective was overweening with ambition, but owing to pragmatic considerations board games have often been used as a proving ground for AI techniques. The ability to play board games with skill is a hallmark of intelligence, and, crucially, board games have precise rules that can be encoded into a computational framework.

> [I]n 2016, a computer program AlphaGo developed by DeepMind, a company owned by Google, created a sensation by defeating the reigning world Go champion...

In a celebrated encounter between man and machine in 1997, IBM's Deep Blue inflicted a shocking defeat of the then world chess champion, Gary Kasparov. Given the storied status of chess as a cerebral game, Kasparov's defeat was unnerving as it heralded the breach of a frontier. At the time many believed that building a machine capable of defeating the world champion at the board game of Go to be a rather distant or unlikely prospect. This belief was based on the fact that Go is played on a much larger board than chess and the number of possible sequences of moves in Go are very much larger than those of chess. But in 2016, a computer program AlphaGo developed by DeepMind, a company owned by Google, created a sensation by defeating the reigning world Go champion, Lee Sedol. As in 1997, this victory was hailed by breathless commentary as inaugurating a new era where the machine was on the verge of overtaking human intelligence. It was nothing of the sort. AlphaGo was a sophisticated tool but it did not approach intelligence by any measure. The software could select the most desirable move at any stage of the game, but had no intrinsic conception of what it was doing or why.

One of the key lessons of the advances in AI is that a machine does not have to be intelligent in the manner of sentient beings to be endowed with abilities that were hitherto the preserve of humans. A highly simplified non-AI example is the case of arithmetic computation. Throughout history, it was a difficult task to carry out calculations such as multiplying two large numbers. There was much human effort expended towards these tasks, including the painstaking creation of logarithm tables. For many decades now, even the simplest of computers can efficiently and reliably carry out such calculations. Similarly, any number of tasks carried out by humans involving routine operations are likely to be solvable using AI.

> *AI is also beginning to make inroads into research areas that involved domain knowledge in science and engineering. One area of universal relevance is healthcare...*

With the extraordinary growth in computational power and availability of volumes of data, contemporary AI promises to extend the above metaphor beyond simple, routine tasks to more sophisticated realms. Many AI tools are already deployed and used by millions. But AI is beginning to make inroads into research areas that involved domain knowledge in science and engineering. One area of universal relevance is healthcare, with AI tools ranging from the monitoring of the health indicators of an individual, providing medical diagnosis based on clinical measurements, to analysis of large-scale population studies. In more esoteric realms, there have been recent developments in using AI for solving highly complex problems such as protein folding or modelling flow dynamics in fluid mechanics. In all instances, it is believed that such advances will result in numerous real-world applications.

History

Much of the early AI work had focused on symbolic reasoning—laying out a set of propositions and making logical deductions from them. But this enterprise quickly ran into rough weather as it was impossible to enumerate all the operational rules in a given problem context. An alternate competing paradigm is connectionism that seeks to overcome the difficulty of explicit descriptions of rules by *learning* them implicitly from data. Loosely inspired by the properties of neurons and their connectivity in the brain, this approach builds *artificial neural networks* that learn the strength (weight) of the connections between neurons.

At different points in time, the success of one paradigm or the other has led to pronouncements by leading figures claiming the imminent arrival of a definitive solution to the problem of computational intelligence. While there were advances made, the challenges proved far more intractable and the hype was typically followed by periods of profound disillusionment leading to a severe reduction of funding available to American academics—a so called AI winter. It is therefore par for the course, when buoyed by its recent success, DeepMind claims that its approaches “could help society find answers to some of the world’s most pressing and fundamental scientific challenges.” To go beyond such claims, the reader interested in the key ideas in AI, the history of the discipline and its boom-bust cycles would profit from two recent popular expositions written by long-time researchers—Melanie Mitchell's *Artificial Intelligence: A Guide for Thinking Humans* (Pelican Books, 2019) and Michael Wooldridge's *The Road to Conscious Machines: The Story of AI* (Pelican Books, 2020).[1]

The question of intelligence

Since its inception AI has struggled with two related problems of profound significance. While defeating world champions at their game is an impressive accomplishment, the real world is a much more ambiguous and messy place to operate in, compared with the universe of games defined by iron-clad rules. Consequently, the successful methods of AI developed to solve narrowly defined problems do not generalise to other challenges involving different aspects of intelligence. To take one instance, robotics research is yet to develop an important skill that a child learns effortlessly – the use of one's hands for delicate tasks. Thus, while AlphaGo churned out the winning moves, the seemingly mundane act of physically repositioning the stones on the board had to be carried out by its human representative. This is not a mere technical detail, for intelligence is not defined by a single skill such as winning at board games. Amongst other things, it encompasses essential aspects of embodied behaviour such as the ability to interact with the environment.

> *[R]obotics research is yet to develop an important skill that a child learns effortlessly – the use of one's hands for delicate tasks.*

More significantly, beyond the technical limitations of AI tools is the profound and bedeviling question of defining intelligence itself.[2] Many AI researchers often presume that the methods developed to solve narrowly defined problems – such as winning at Go – would help leapfrog into the realm of general intelligence. This rather brash belief has been met with scepticism, from both within the community and in a more hostile manner from those steeped in older disciplines such as philosophy and psychology. The key bone of contention has been whether intelligence was amenable to being substantially or fully captured in a computational paradigm or had an ineffable, irreducible human core. The sentiment of derision and hostility to the claims of AI in some quarters is reflected in a well-known and early attack by the philosopher Hubert Dreyfus who titled his 1965 report, *Alchemy and Artificial Intelligence*. In response, a well-known AI researcher called Dreyfus's views 'a budget of fallacies'.

On the other extreme is a view of unbridled optimism about AI's ability to break all barriers thereby transcending the limits of biology, a notion known as Singularity. Propounded by the 'futurist' Ray Kurzweil, this view holds that the capabilities of AI systems would compound exponentially leading to an explosion of machine intelligence outstripping the capabilities of human minds. Despite the ridiculousness of Kurzweil's argument based on exponential growth in technology, it has attracted a camp of fervid followers. The idea of a Singularity lacks serious intellectual grounds and is best seen as a form of 'technological Rapture'.[3]

> *The key bone of contention has been whether intelligence was amenable to being substantially or fully captured in a computational paradigm or had an ineffable, irreducible human core.*

An AI researcher who does not shy away from defining intelligence is Stuart Russell, the first author of *the* most popular textbook on AI. In *Human Compatible: AI and the Problem of Control* (Allen Lane, 2019), Russell defines humans as "intelligent to the extent that our actions can be expected to achieve our objectives" (Russell, *Human Compatible*, 9). The definition of machine intelligence follows the same lines. Such a crisp definition does pin down the elusive notion of intelligence, but as anyone who has contemplated the notion of utility in economics would recognise, it does so by shifting the burden of meaning to a precise description of our objectives. Unlike Mitchell and Wooldridge, Russell is stylistically terse and expects a high level of engagement and yields no quarter to his reader. While hugely thought-provoking, *Human Compatible* is also an idiosyncratic narrative that hops from lucid arguments to abstruse conjectures and conclusions.

> *[A] recent study found that none of the hundreds of AI tools developed for detecting Covid were effective.*

Human Compatible is also markedly different from other AI expositions in its central concern—the dangers of future AI outstripping human intelligence. Without invoking Terminator-like dystopian Hollywood imagery, Russell argues that in the future AI agents could combine in unintended and harmful ways. Warning against the belief that such an eventuality was highly unlikely or impossible, Russell points to the story of the physicist Leo Szilard figuring out the physics of a nuclear chain reaction as he was annoyed by the stated belief of Ernest Rutherford that the idea of atomic power was moonshine. The horrors of atomic warfare followed. Much of *Human Compatible* is given over to thinking through on how to guard against such a possibility with AI. Here, we note that Wooldridge is unconvinced by this argument. For him, the experience of six decades of AI research suggests that "human-level" AI is unlike the "simple mechanism" of a nuclear chain reaction (Wooldridge, *The Road to Consciousness*, 243).

AI in the world

Philosophical debates on the nature of intelligence and the fate of mankind are enriching but ultimately undecidable. In reality, AI research runs along two largely distinct tracks – as cognitive science and as engineering – wherein most researchers are concerned with specific problems and are often indifferent to the larger debates. However, in the public discourse the objectives and claims of these two approaches are often conflated, resulting in much confusion. Pertinently, within the discipline, terms such as neurons and learning have specific mathematical meaning. But for the lay person they immediately evoke their commonsense connotation leading to a serious misunderstanding of the entire enterprise. To be clear, a neural network is analogous to the human brain in only a highly superficial sense, and learning is a broad set of statistical ideas and methods that are effectively sophisticated forms of curve fitting or decision rules.

> *Since the breakthroughs almost a decade ago, deep learning has swept across many disciplines and has almost completely replaced other methods of machine learning.*

The idea of neural networks that learn from data appeared some decades ago but was largely abandoned and deemed ineffective. In 2012, neural networks garnered renewed academic and commercial interest with the development of *deep learning* which led to substantial improvements in methods for image recognition and speech processing.[4] The current wave of successful AI methods, including AlphaGo and its successors and widely used tools such as Google Translate, use deep learning where the adjective does not signify profundity but the prosaic fact that the networks have many layers. Since the breakthroughs almost a decade ago, deep learning has swept across many disciplines and has almost completely replaced other methods of machine learning. This paradigmatic dominance received official anointment when in 2018 three of its pioneers were decorated with the highest honour in computer science, the Turing Award.

It is an iron law of AI that success is followed by hype and hubris. Thus, as if on cue, one of the Turing triumvirate, Geoff Hinton proclaimed in 2016: "We should stop training radiologists now, it's just completely obvious within five years deep learning is going to do better than radiologists." The failure to deliver us from flawed radiologists and many other problems with the method did not deter Hinton from claiming in 2020 that "deep learning is going to be able to do everything". Incidentally, a recent study found that none of the hundreds of AI tools developed for detecting Covid were effective.

> *It is an iron law of AI that success is followed by hype and hubris.*

To look beyond hyperbole and understand the nature of the problems and concerns that arise with contemporary learning-based AI tools, we need to look at how they are created. Consider the hypothetical task of detecting chairs in an image. One may think of ways to detect elements of a chair: legs, arm and back rests, cushions etc. It will be obvious that a bewildering range of combinations of such elements are possible, all of which would be recognisable as a chair. And then there are significant exceptions such as a bean bag that can trump any rules we may define for the constituents of a chair. It is precisely the limitations of such symbolic, rule-based deduction that methods like deep learning seek to overcome. Instead of trying (and failing) to define rules that encompass the entire gamut of chairs, we may instead collect a number of images of chairs and other objects and provide them to a neural network as *inputs*, along with the correct answers (*output* of chair vs. non-chair). In a *training phase*, a deep learning method would then modify the weights of the connections in the network so as to mimic the desired input-output relationships as best as possible. If done well, the network should now be endowed with the ability to answer the question of whether new, previously unseen *test* images contain chairs. It stands to reason that building such a chair-recogniser would require a very large number and variety of images of chairs. We may now extend the analogy to any number of categories one can imagine: chairs, tables, trees, people etc. all of which appear in the world in their glorious and maddening diversity. The concomitant need for suitably representative images of objects reaches staggering proportions.

> *While deep learning methods have worked surprising well, they are often unpredictable and unreliable in their behaviour.*

Deep learning demonstrated significant advances in image recognition in 2012 precisely because of the conjunction of relatively cheap and powerful hardware, and the explosive growth of the internet that enabled researchers to build a large-scale dataset, known as ImageNet, contained millions of images labelled with thousands of categories.

While deep learning methods have worked surprising well, they are often unpredictable and unreliable in their behaviour. For instance, it is well known that small changes to images that are imperceptible to the human eye can lead to wildly incorrect classification of images, e.g., an American school bus being labelled as an ostrich. It is also recognised that in other instances correct results can arise out of spurious and unreliable statistical correlations and not out of any real understanding. For instance, a boat is correctly recognised only when it is surrounded by water in the image. In other words, the method has no model or conception of what a boat might look like. In the past, such problems and limitations of AI methods might have remained matters of academic concern. However, this time it is different as a number of AI tools have been plucked out of the laboratory and deployed into the real world, often with serious and harmful consequences.

Bias and fairness

Even before the current wave of interest in deep learning methods, owing to a relentless thrust towards automation, a number of data-driven methods have been commercially developed and deployed, notably in the United States but increasingly across the world including in India. One such example that has achieved great notoriety is COMPAS, a tool used in US courts to provide a score for recidivism risks to help a judge decide on the duration of the sentence to be awarded. Such a tool uses statistics from existing criminal records to determine the likelihood that a defendant would commit a crime if released early. A well-known investigation has demonstrated that even without explicitly setting out to do so, the tool was significantly biased against black people. Judges relying on machine learning to determine the duration of the sentence end up discriminating on the basis of race. Of even greater vintage is the use of fingerprints and face images for biometric identification and authentication. Owing to their utility in surveillance and forensic analysis, face recognition tools have been rapidly adopted by many law enforcement and other state agencies. Dubious methods of so-called emotion recognition and other forms of 'affective computing' have also been deployed in a number of contexts including hiring of employees as well as newer and more intrusive forms of surveillance.

In the US, some important studies have shown that many of the commercially available face recognition tools are deeply flawed and discriminate against people of colour. One audit of commercial tools has shown the difference in face recognition error rates between white men and black women can be as high as 35%, leading to growing calls for the halt of their deployment. The implications of face and emotion recognition for human rights and welfare – around the world especially in authoritarian China, and increasingly in India – are extremely serious and warrant a fuller consideration than is possible in this essay.

The sources of the inherent problems with decision making based on data extracted from the real world are numerous, many of which are grouped under the rubric of bias.[5] A major source of bias in the case of face recognition is that many datasets used to build the tools had far fewer examples of people of colour. Another problem is that the past is a poor normative guide for the contours of the society we would wish to build. But, as in the case of recidivism modelling in the US, if an AI method relied on past records it would disproportionately penalise people from poorer communities who have historically experienced discrimination in the form of higher incarceration rates. Similarly, if one were to hypothetically build a tool for automating hiring people in India for say a professional position, models based on past hires would automatically lead to a caste bias, even if caste was not used as an explicit factor in the decision. The broad argument here is captured in the title of Cathy O'Neil's popular book, *Weapons of Math Destruction: How Big Data Increasing Inequality and Threatens Democracy* (Penguin Books, 2016) which details a number of such instances drawn from the American experience.

> *[I]f one were to hypothetically build a tool for automating hiring people in India for, say, a professional position, models based on past hires would automatically lead to a caste bias...*

In almost all cases, AI methods do not learn from the world, but from a proxy in the form of a dataset. For long, academic AI research has paid little attention to the design of data collection or AI methodology, and ethical oversight has been non-existent. As in the case of the poor accuracy of face recognition tools, it has taken a significant amount of effort by scholars from a range of disciplines to create the discussion of bias in AI tools and datasets, and their ramifications for society, especially amongst the poor and traditionally discriminated communities. Apart from bias, another area of concern is the virtual impossibility to reason about the decisions of most contemporary AI tools or interpret their results. This problem of 'explainability' has serious implications for transparency, especially in the legal sense, since those affected by a decision often have the right to be given the reasoning used to arrive at it.

More broadly, there has been some interest within the computer science community to examine these problems in a formal sense, leading to a growing body of research work including academic conferences and an online textbook in the making. One key outcome of this exercise has been a theoretical understanding that multiple notions of fairness cannot all be simultaneously satisfied, resulting in impossibility-of-fairness theorems. This implies that AI researchers and practitioners need to recognise the trade-offs involved in design choices and understand their societal implications. But, as we shall see in the second part of this essay, important as these considerations are, they are seldom adequate as the rapid transition of contemporary AI from the research laboratory into the public domain has unleashed a wide range of consequences.

This is the first part of this essay. Part-II can be read here.

The views expressed in this article are personal.

Footnotes:

[1] Mitchell's account carefully delineates the different approaches to AI, while methodically taking apart the inflated claims. Wooldridge's provides a more detailed and chronological account from the perspective of an insider and fleshes out the rather contentious nature of debates within the AI community.

[2] Here we note that the popular Turing test merely sidesteps the definitional question and translates it into one of replication of intelligent behaviour. Such a test tells us nothing about the attributes of intelligence or the means of achieving them.

[3] Andrian Kreye quoted in Mitchell, *Artificial Intelligence*, 51. On the paradigm of exponential compounding in natural and social phenomena, see a recent clinical dissection, Vaclav Smil, *Growth: From Microorganisms to Megacities*, Cambridge: The MIT Press, 2019.

[4] Deep learning is the latest method in an entire sub-field known as *machine learning* which is built around principles of statistical inference and largely focuses on optimisation techniques. While machine learning practitioners see their work as different from conventional AI, this distinction is lost in popular perception.

[5] Bias is yet another ambiguous term in AI that has a number of different meanings. To add to the confusion, for most machine learning researchers, the natural meaning of bias would be statistical bias which has a specific mathematical definition in the context of estimation.