LRB

Vol. 43 No. 2 · 21 January 2021

Insanely Complicated, Hopelessly Inadequate

Paul Taylor

The Promise of Artificial Intelligence: Reckoning and Judgment
by Brian Cantwell Smith.
MIT, 157 pp., £20, October 2019, 978 0 262 04304 5

Rebooting AI: Building Artificial Intelligence We Can Trust
by Gary Marcus and Ernest Davis.
Ballantine, 304 pp., £22.50, September 2019, 978 1 5247 4825 8

The Book of Why: The New Science of Cause and Effect
by Judea Pearl and Dana Mackenzie.
Penguin, 418 pp., £10.99, May 2019, 978 0 14 198241 0

When I first studied artificial intelligence in the 1980s, my lecturers assumed that the most important property of intelligence was the ability to reason, and that to program a computer to perform intelligently you would have to enable it to apply logic to large bodies of facts. Logic is used to make inferences. If you have a general rule, such as ‘All men are mortal,’ and a specific fact, ‘Socrates is a man,’ you, or your computer, can deduce that Socrates is mortal. But it turns out that many of the problems we want intelligent computers to help us with can’t straightforwardly be solved with logic. Some of them – the ability to recognise faces, for example – don’t involve this kind of reasoning. In other cases – the diagnosis of disease would be an example from my own field – the difficulty lies in how to describe the concepts that the rules and facts express. The problem is often seen as a matter of how to standardise terminology. If you want a doctor’s computer to use rules to infer what is wrong with a patient, these rules must be expressed using the same words as the ones used in the patient’s records to describe their symptoms. Huge efforts are made to constrain the vocabulary used in clinicians’ computer systems, but the problem goes deeper than that. It isn’t that we can’t agree on the words: it’s that there aren’t always well-defined concepts to which the words can be attached. In *The Promise of Artificial Intelligence*, Brian Cantwell Smith tried to explain this by comparing a map of the islands in Georgian Bay in Ontario with an aerial photograph showing the islands along with the underwater topography. On the map, the islands are clearly delineated; in the photograph it’s much harder to say where each island ends and the sea begins, or even exactly how many islands there are. There is a difference between the world as we perceive it, divided into separate objects, and the messier reality. We can use logic to reason about the world as described on the map, but the challenge for AI is how to build the map from the information in the photograph.

Cantwell Smith argues that if we seem to inhabit a world that is constructed of ‘discrete, well-defined mesoscale objects exemplifying properties and standing in unambiguous relations’, that is an achievement of our intelligence, not a truth that can be used when engineering an artificial intelligence. This will resonate with anyone who has tried to express seemingly straightforward concepts in sets of rules, only to be defeated by the complexity of real life. One of the standard terminologies used for the computerisation of medical records includes arteries as a subclass of soft tissue, which seems not unreasonable if 'soft tissue' is taken to include anything that isn't bone, but has the consequence that aortic aneurysm is classified as a disorder of soft tissue, which may be logically correct but feels out of place. The problem is that any attempt to devise a scheme that is rigorously logical inevitably diverges from the way we actually talk about the world. The point, and the lesson I take from Cantwell Smith's book, is that we can navigate these contradictions because we understand that our ideas are imperfect.

One of the reasons Cantwell Smith believes that our oversimplified assumptions about ontology can explain the failure of the symbolic logic employed by what's sometimes called Good Old-Fashioned AI, or GOFAI, is that more recent, and more successful, approaches to AI don't depend on this kind of symbolic reasoning. Over the last forty years the extraordinary increase in the rate of accumulation of digital data and the equally dramatic drop in the price of processing power have made possible purely data-driven approaches to machine learning. These systems make predictions based on correlations observed among vast quantities of data. They break calculations down into billions of simpler ones, and learn by iteration, altering the weight given to each piece of information at each stage, until the output of the entire network of calculations conforms to a predetermined target. Artificial neural networks have proved spectacularly successful at tasks such as generating captions for images, recognising spoken words and identifying winning moves in chess. Cantwell Smith takes the contrast between the relative success of this kind of machine learning and the relative failure of GOFAI as evidence that Descartes was wrong to say that understanding must be grounded on 'clear and distinct' ideas. It is, he argues, the capacity to represent an ineffable reality that exists at a 'subconceptual' level which underpins intelligence.

Given the extent of the paradigm shift in AI research since 1980, you might think the debate about how to achieve AI had been comprehensively settled in favour of machine learning. But although its algorithms can master specific tasks, they haven't yet shown anything that approaches the flexibility of human intelligence. It's worth asking whether there are limits to what machine learning will be capable of, and whether there is something about the way humans think that is essential to real intelligence and not amenable to the kind of computation performed by artificial neural networks. The cognitive scientist Gary Marcus is among the most prominent critics of machine learning as an approach to AI. *Rebooting AI*, written with Ernest Davis, is a rallying cry to those who still believe in the old religion.

Marcus and Davis maintain that, despite the current excitement around machine learning, it will always fall short of achieving a real general intelligence. Such intelligence, they argue, requires the use of symbolic logic. Where Cantwell Smith's argument is best illustrated with examples in which computers are programmed to make sense of complex visual scenes (such as aerial photographs of partially submerged islands), Marcus and Davis focus on attempts to understand language. They argue that GOFAI is needed if computers are to understand what we're telling them. Much of the difficulty with the application of machine learning to language is that it is, in practice, the application of machine learning to statistics about language, not to language itself. You can learn something about the meaning of a word by asking which other words are commonly used in conjunction with it. If we apply machine learning, words with similar meanings can be grouped together, so that typing 'car' into a search engine will retrieve documents indexed under 'automobile'.

It is tempting to assume that these statistics can function as a representation of meaning, and models built using this approach have proved remarkably successful. In 2017 Google published an algorithm called ‘Transformer’ which, instead of processing the words in a sentence one at a time, as humans do, analyses them simultaneously. By exploiting the potential for parallel processing in massive networks of computers, the algorithm can be trained on larger and larger collections of texts. Transformer also lies behind the recent success of Google DeepMind’s AlphaFold, the program which, it was announced in November, has solved one of the big challenges in biomedicine, predicting how the long chains of molecules in proteins will fold into the complex 3D shapes that determine their function. There is something pleasing in the idea that a technique designed to extract meaning from a sequence of symbols can be used to work out the biological significance of a molecule from its chemical composition.

Last year, Google’s work on natural language processing was the subject of a piece co-written by Timnit Gebru, one of the leaders of its ‘ethical AI’ team. The article expressed concerns about the work’s carbon footprint – the extraordinary scale of computation involved means that the carbon dioxide emitted in training Transformer is equivalent to 288 transatlantic flights – and about the way it looks at language. Because it is trained on text that Google harvests from the internet, its calculations reflect the way language has been used in the past or is used now. The problem isn’t just that its outputs therefore reflect our biases and prejudices, but that they crystallise them and, because the programs are inscrutable, conceal them. The paper also discusses the opportunity cost involved in pursuing this approach to the exclusion of others, including those advocated by Marcus and Davis. Google’s response was to shoot the messenger, sacking Gebru and then claiming she had resigned. Given that one very dangerous aspect of AI is that it amplifies the already extraordinary power of a very small number of massive corporations, this authoritarian behaviour is alarming. On the other hand, one of its immediate effects has been to galvanise workers at Google into forming a trade union.

It has proved difficult to use statistical models to assemble the meaning of longer passages of text from their component parts. Perhaps, as Marcus and Davis argue, in order to understand the meaning of the word ‘car’, you have to know not just how the word ‘car’ is used but also something about cars. They believe it’s essential to apply symbolic reasoning to the ill-defined concepts we think and talk about, however hard it is to program computers to do that. When we understand what other people tell us, we aren’t just hearing what they’re saying, we’re also drawing on knowledge of the way the world works, the way people and other objects behave, knowledge that is so obvious to us that we aren’t conscious of using it.

One of the most heroic endeavours of the GOFAI era was a project called ‘Cyc’, which set out to express the entirety of this kind of common sense in a computable form. A paper from 2006 explaining its approach to ontology uses the example of Martin Luther’s burning of a papal bull, and shows how information about this would be represented. It would be necessary to make a distinction between the social event, attended by Luther and his followers, and the combustion event, which involved the destruction not just of a physical object – the papal bull – but also of a conceptual formation, papal authority. The meticulous conceptual engineering involved seems both insanely complicated and hopelessly inadequate. The project began in 1984 and, after more than a thousand man years of effort, it has now amassed more than 24 million facts. But it doesn’t, as Marcus and Davis concede, seem to have enabled the hoped for advance in machine understanding. It also seems dated in a different way. The kind of common sense Cyc aimed to identify and represent is the knowledge that leads us to interpret the statement ‘the police arrested the protesters because they feared violence’ differently from ‘the police arrested the protesters because they advocated violence.’ Even if it were possible to identify and record a set of suitable ‘facts’, would these facts be universally agreed on? It’s hard to accept Marcus and Davis’s insistence that, despite decades of failure, symbolic reasoning is the only way forward. They admit they can’t explain how such approaches can be made to work, suggesting only that researchers focus not on accumulating large collections of facts, but on thinking carefully about the key concepts that underpin our understanding of the world.

The concept of causality is central to this debate because we are active participants in the world as computers are not. We observe the consequences of our interventions and, from an early age, understand the world in terms of causes and effects. Machine learning algorithms observe correlations among the data provided to them, and can make astonishingly accurate predictions, but they don't learn causal models and they struggle to distinguish between coincidences and general laws. The question of how to infer causality from observations is, however, an issue not just for AI, but for every science, and social science, that seeks to make inferences from observational rather than experimental data.

This is a question that Judea Pearl has been working on for more than thirty years. During this time, he and his students have, as Dominic Cummings's eccentric Downing Street job advert put it, 'transformed the field'. In the 1980s, it seemed to some researchers, including Pearl, that because one characteristic of intelligence was the ability to deal with uncertainty, some of the problems that couldn't be tackled with logic could possibly be solved using probability. But when it comes to combining large numbers of facts, probability has one huge weakness compared to logic. In logic, complex statements are made up of simpler ones which can be independently proved or disproved. It is harder to deal with complex probabilities. You can't work out the probability of someone having both heart disease and diabetes from the separate probabilities of their having diabetes or heart disease: you need to know how the likelihood of having one affects the likelihood of having the other. This quantity – the probability of something happening given that something else has already happened – is known as a conditional probability. The main difficulty in using probability is that even a modest increase in the number of concepts to be considered generates an explosive increase in the number of conditional probabilities required.

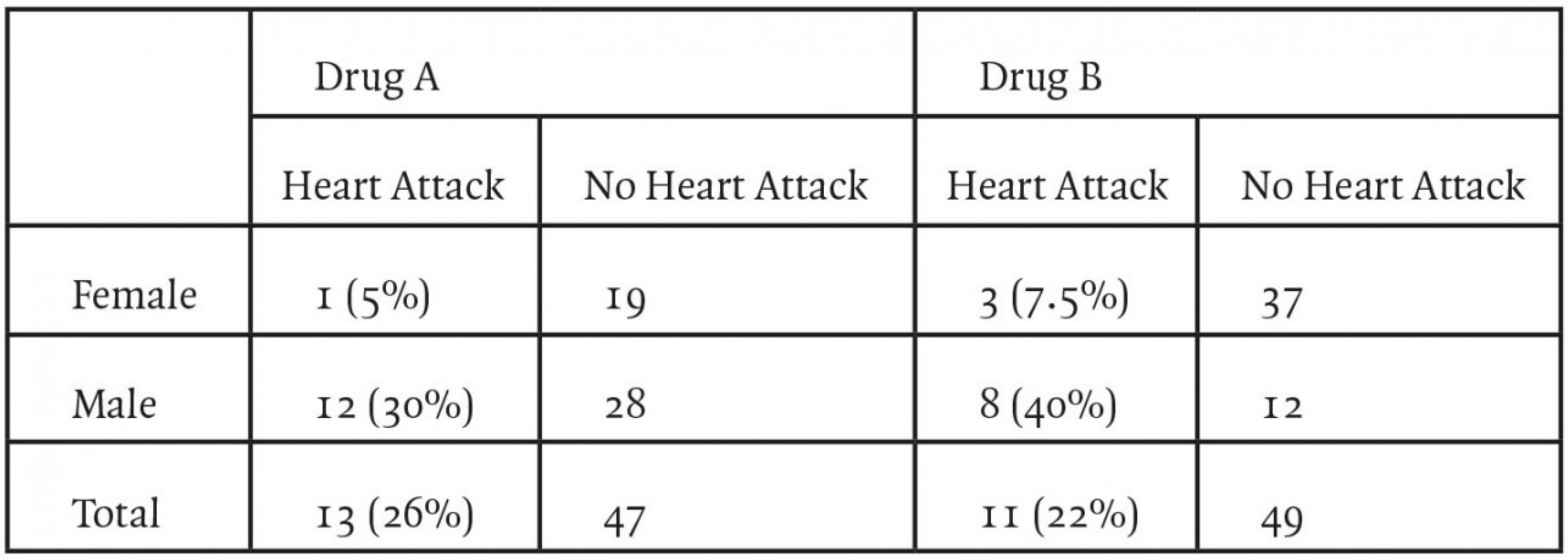
One way out of this problem builds on Bayes' Theorem, which the Reverend Bayes is said to have derived to win an argument with David Hume about the value of eyewitness testimony as evidence of miracles. The theorem tells us that the probability of a hypothesis being true, given the evidence we have in support of it, depends on two quantities. One is the quality of the evidence. The other is the prior probability of the hypothesis: how plausible it would seem in the absence of evidence. The prior probability is fundamental to Bayesian thinking: a non-Bayesian statistician given a set of data would draw conclusions using only that data, but a Bayesian statistician considers his or her prior beliefs and tries to work out how they can best be updated with the additional information provided by the data. This is novel because the mathematics of probability is being applied to a measure of confidence in a belief or hypothesis rather than a measure of chance, such as the odds of rolling a dice and getting a six, and because of its requirement that new information be interpreted in the context of what is already known.

Pearl's first contribution to AI was to use Bayesian ideas to avoid the need for a huge number of conditional probabilities. Inspired by early work on neural networks, he started to draw simple diagrams of the connections between the variables in the problems he was trying to solve. Each variable (heart disease, obesity, diet) was allotted one or more states (present or absent, true or false, good or bad) and arrows were used to connect pairs of variables when a change in the state of one would be likely to lead to a change of state in the other: an arrow from obesity, for example, pointing towards diabetes. You have to know, or be willing to estimate, the prior probabilities of each state for every variable, but you only need to know the conditional probabilities for the variables connected by arrows. These diagrams encode our prior assumptions, and massively reduce the number of conditional probabilities required, which makes it easier to calculate the probability of any combination of states and to update it as more information becomes available.

We might think of these diagrams, which became known as Bayesian networks, as maps showing the causal connections between variables. In fact the connections in the diagrams needn't necessarily be causal, and at first Pearl was careful not to talk about causality, but it's hard to avoid the word: the arrows have a direction and so represent something more than a correlation – they suggest that one concept influences the other. But eventually Pearl did start referring to them as causal diagrams. In *The Book of Why* he argues that epidemiologists, economists and other social scientists have taken the maxim 'correlation does not imply causation' too much to heart and have, as a result, spurned opportunities to use data to understand the world in terms of causes and effects.

By 1950 large research studies had repeatedly demonstrated that people who smoke were much more likely to get lung cancer than those who didn't, and that the more they smoked, the more likely it became. But many people – including some prominent statisticians – argued that this evidence of association could not be taken as evidence of causation. The key difficulty is that while we can measure the proportion of smokers who develop lung cancer, that's not the same thing as being able to measure the proportion of smokers who develop lung cancer because they smoke. Maybe, a tobacco lobbyist could argue, people who smoke also like to drink, and it is the alcohol that causes cancer. Statisticians call the intrusion of another variable to produce a spurious association 'confounding'. To eliminate alcohol as a possible confounder would require using sub-groups with different levels of alcohol consumption and assessing the danger of smoking in each one, thereby 'controlling' for the confounding variable. It's tempting to control for the widest possible range of confounders, but it isn't always the right thing to do.

Pearl demonstrates this with an example based on Simpson's paradox. Imagine we are examining the efficacy of two different drugs used to reduce the risk of heart attack, both of which work by lowering blood pressure. Drug A is more effective than Drug B at preventing heart attacks in women and in men, but paradoxically it's less effective overall. If we control for gender, we will prefer drug A; if we don't, we will prefer drug B. But should we?



	Drug A		Drug B	
	Heart Attack	No Heart Attack	Heart Attack	No Heart Attack
Female	1 (5%)	19	3 (7.5%)	37
Male	12 (30%)	28	8 (40%)	12
Total	13 (26%)	47	11 (22%)	49

There is one clear gender effect here: drug A is used by more men than women, drug B by more women than men. It might seem obvious that gender is determining the choice of drug (and that gender is therefore a confounder for which we should control). But if we replace the labels 'men' and 'women' with 'high blood pressure' and 'low blood pressure', we can't make the same argument because we don't know whether the level of blood pressure is a cause or an effect of the choice of drug. Pearl concludes that our understanding of the underlying causal processes tells us how to analyse the data. We never start from scratch: before we can begin to estimate the causal impact of one variable on another, we must have a model of the domain and some understanding of the roles other variables play.

The richer the models, the more sophisticated the reasoning we can deploy. Courts address questions of causality using the ‘but for’ test. If, but for the fact you were speeding, I would not have sustained these injuries, it follows that you caused them. This test involves counterfactual reasoning: we have to imagine the world as it would be if you hadn’t been speeding. We can work out the cause of something from data using the same process, comparing an observed quantity (the number of smokers who died of lung cancer) with an estimate of the counterfactual quantity (the number who would have died had they not smoked). Consider a study that found there were more deaths among obese people, equivalent to a hundred thousand excess deaths per annum. It is tempting to say that obesity is the cause of these excess deaths. But we don’t get an accurate estimate of the causal impact of obesity by comparing the mortality rate in the obese population with the rate in the non-obese population. Instead we need to compare it to the mortality rate in a counterfactual world where the obese are not obese. The estimate will depend, crucially, on what we think might explain the difference: did they exercise more, did they have a better diet? Some insist that we can only make causal inferences about the impact of imaginable interventions: we can’t estimate the number of deaths caused by obesity, only the number of lives that could be saved by, for example, advocating a healthier diet.

It follows that we can’t estimate the consequences of any causal factor for which there isn’t a straightforwardly effective intervention, like sexism or racism. Pearl regards this as an unnecessary restriction since he sees causal diagrams as models of our assumptions about the world, more like GOFAI’s abstractions than neural networks learned from data.

The causal models described in *The Book of Why* are typical of work in data science: they describe relationships between what are assumed to be well-defined concepts about which we can collect reliable data. They help formalise hypotheses and inform analysis. They don’t help bridge the gap between the data and the real world. Where the concepts are well defined, as in the link between smoking and cancer, the data may be sufficiently robust to answer causal questions, but in many areas things are less clear. As a contribution to AI, rather than data science, Pearl’s approach seems to occupy a halfway house between GOFAI and machine learning, but it exists in the ontological world of objects, properties and relations, the world humans create, and lacks the ability of machine learning to make direct use of data about the real world.

Although machine learning seems to dominate the field, some Bayesian networks are currently in use. Babylon Health, an app that provides advice to patients thinking of consulting a GP, uses a model based on 500 million concepts. Patients using the app interact with a chatbot that asks about their symptoms. The system contains estimates of the prior probabilities of different illnesses, and the conditional probabilities of those illnesses given particular symptoms. It uses these to generate estimates of the probability of each explanation for a patient's symptoms and to advise them whether to seek help. It is hard to imagine building a system on this scale from raw data as a machine learning system would do – although as collections of data get larger the idea seems less implausible – and hard to believe that ignoring the medical profession's accumulated knowledge of symptoms and diseases would be a sensible approach. Although it uses doctors' knowledge, Babylon is unpopular with GPs, who dislike the business model and the fact that it was brought to market without the kind of rigorous evaluation required of new drugs. Many GPs gleefully tweet its mistakes: a breast lump diagnosed as osteoporosis, a 67-year-old smoker with sudden onset chest pain told they probably have gastritis. The system clearly makes errors; it's not clear whether it makes more than we should allow. Last year a Babylon research team published a paper describing a Bayesian network, presumably similar to the one used in its commercial products, which contained information on the conditional probabilities of symptoms given diseases, and of diseases given risk factors. They found that if you enter symptoms that are highly correlated with multiple diseases sharing a common risk factor, this increases the network's estimate of the probability that the risk factor is present, which can lead it to suggest diagnoses that are highly correlated with the risk factor, even when it isn't a plausible explanation of the symptoms: the network might, for example, infer that a patient with chest pain could be obese and then suggest diabetes as a diagnosis. The team found that using a form of counterfactual reasoning – asking, for example, for each possible diagnosis, how the array of symptoms would change if the diagnosed condition were cured – significantly improved the accuracy of the network's diagnoses, especially in cases doctors found difficult.

It still seems likely that future advances in AI will come from neural networks, simply because of the sheer scale of research now devoted to them. Yoshua Bengio, one of the pioneers of neural networks, argues that the instability of a network can serve as a test of the accuracy of causal assumptions. Perhaps, he says, a network presented with evidence of a correlation could guess the direction of the causal arrows and then measure 'regret' – how much experience is required before a learning algorithm recognises it needs to do things differently – to assess whether it was right. The hope is that data-driven machine learning will be able to move beyond simple pattern-recognition and start to develop the organising theories about the world that seem to be an essential component of intelligence. AI software has, so far, only managed this in very constrained environments, with games such as Go or chess.

Cantwell Smith argues that the key challenge facing all attempts to build AI, either by classical knowledge engineering or contemporary machine learning, is philosophically similar to what Bengio was getting at: the recognition that our ideas about the world are inevitably incomplete. Humans know that there is a world, that their ideas are representations of the world, and that these representations must defer to a world which will always, to some degree, elude representation. Although machine learning systems are better than GOFAI at evading the limitations imposed by naive assumptions about ontology, they still lack an awareness of their relationship to the world. Requiring them to develop this might be too exacting a demand. A computer can play chess to superhuman levels and yet have no concept of what chess is, what place chess has in the world, or even that there is a world. Does this mean that its behaviour isn't intelligent? Perhaps there is a limit to what a computer can do without knowing that it is manipulating imperfect representations of an external reality. To take one pressing example, Cantwell Smith argues that safely controlling a self-driving car in an urban environment will require the kind of judgment that makes such awareness necessary. Perhaps, but it seems at least possible that careful engineering could make a car that would be safe enough, even if it doesn't really know what it is doing.