Article

Not peer-reviewed version

Deep Learning and Knowledge

Donald Angus Gillies *

Posted Date: 16 October 2024

doi: 10.20944/preprints202410.1221.v1

Keywords: Deep Learning; Opaqueness; Knowledge How; Knowledge with a Rationale

 doi:10.20944/preprints202410.1221.v1

Article

Deep Learning and Knowledge

Donald Angus Gillies

University College London; donald.gillies@ucl.ac.uk

Abstract: This paper considers the question of what kind of knowledge is produced by deep learning. Ryle's concept of *knowledge how* is examined and is contrasted with *knowledge with a rationale*. It is then argued that deep neural networks do produce knowledge how, but, because of their opacity, they do not in general, though there may be some special cases to the contrary, produce knowledge with a rationale. It is concluded that the distinction between knowledge how and knowledge with a rationale is a useful one for judging whether a particular application of deep learning AI is appropriate.

Keywords: Deep Learning; Opaqueness; Knowledge How; Knowledge with a Rationale

1. Introduction

Starting in 2012, there has occurred in AI what is known as *the deep learning revolution*. Deep learning uses neural networks which had a long history of development before 2012. However, deep learning also has a specific feature namely that the artificial neural networks are arranged in layers, and the number of these layers gives the depth of the system. Deep neural networks are trained on large data sets and are then able to make predictions. They thus represent a form of computer induction. An excellent account of the deep learning revolution is given in Sejnowski (2018), or rather we should say that this is an account of the beginning of the revolution which seems to be far from over. Already it has produced remarkable advances, but also worries. Can we always trust these new AI systems or might they sometimes lead us astray? The present paper is an attempt to shed some light on this important question by considering what kind of knowledge is produced by deep learning.

This question has been discussed in an illuminating fashion by Otsuka in section **4.3**, pp. 124 - 143 of his 2023. Otsuka's view is that we can divide knowledge into two kinds and that deep learning produces one kind of knowledge but perhsaps not the other. This is how he puts it (2023, pp. 137-8):

"Ernst Sosa's dichotomy of knowledge will help us to see this contrast. According to Sosa, there are two types or varieties of knowledge: one is what he calls *animal knowledge*, and the other is *reflective knowledge*. … Given this distinction, what kind of knowledge can we attribute to deep learning models? It would be hard to deny that they already have some form of animal knowledge … In contrast, it is not evident whether they also have or will acquire reflective knowledge."

In what follows, we will defend this 'two kinds of knowledge view', but with one difference. Rather than using Sosa's dichotomy of knowledge, we will use a developed and somewhat altered form of Gilbert Ryle's distinction between knowing how and knowing that. We will next consider Ryle's distinction, and later explain how it differs from Sosa's.

2. Knowlege How

Ryle's distinction between knowing how and knowing that is put forward in his (1945-6) paper and his 1949 book *The Concept of Mind*, Chapter II, pp. 25-61. Examples of knowing how are knowing how to speak a language, knowing how to play the piano, knowing how to play chess and so on. These examples can be contrasted with knowing that p, where p is some proposition. Ryle gives the example of knowing that Sussex is an English county (1949, p. 59). Curiously enough, Ryle expounds this distinction in connection with what he calls (1949, p. 25): "that family of concepts

 doi:10.20944/preprints202410.1221.v1

ordinarily surnamed 'intelligence'." However, as he was writing in the 1940s, it is most unlikely that he would have considered artificial intelligence. Yet it turns out that some of the what he says does apply to AI. Ryle was a leading member of the school of ordinary language philosophy, and he gives a list of words used in everyday life which belong to the 'intelligence family'. This list includes the following: clever, acute, witty, cunning, wise, judicious.

In line with his general philosophical approach, Ryle chooses to analyse the concept of witty. Perhaps we should imagine as background a group of Oxford dons making witty conversation at the high table of some ancient college. Ryle says (1949, p. 30):

" … there are many classes of performances in which intelligence is displayed, but the rules or criteria of which are unformulated. The wit, when challenged to cite the maxims, or canons, by which he constructs and appreciates jokes, is unable to answer. He knows how to make good jokes and how to detect bad ones, but he cannot tell us or himself any recipes for them."

This was exactly the problem which AI researchers encountered in the 1970s when they attempted to develop what were known as rule-based expert systems. When AI researchers tried to get the experts to tell them the rules which they used to carry out their skilled tasks, they found that very often the experts were unable to say what rules (if any) they were using. They just knew how to perform the task without any idea of how they managed to do it.[1]

Ryle does, however, think there is a connection between knowing how to do something and some set of rules, principles, maxims or canons. (Ryle is fond of using a variety of different words!) Yet he thinks that in most cases knowing how to do something comes first and it is only after a practical competence at doing something is acquired that the rules, principles, etc. for doing it can be extracted. Thus, he says (1945-6, pp. 11-12):

"Indeed we could not consider principles of method in theory unless we or others already intelligently applied them in practice. Acknowledging the maxims of a practice presupposes knowing how to perform it. … We certainly can, in respect of many practices, like fishing, cooking and reasoning, extract principles from their applications by people who know how to fish, cook and reason. Hence Izaak Walton, Mrs. Beeton and Aristotle."

Ryle then goes on to ask what is the use of the principles extracted from the study of some successful practice. His answer is that they are useful for teaching purposes. As he says (1945-6, p. 12):

"What is the use of such formulae … ? The answer is simple. They are useful pedagogically, namely, in lessons to those who are still learning how to act. They belong to manuals for novices."

Ryle's idea is that the real expert performs a skilled task without calling to mind any rules, formulae etc. associated with it. He or she just does it.

This certainly applies to knowing how to speak a language. Children learn this by listening to adults and receiving occasional corrections when they speak, e.g. "I runned all the way home." "No dear, you *ran* all the way home." Grammatical rules were indeed extracted from some languages in order to teach them, but the language had been spoken for a long time before such grammatical rules were extracted, and, even when such grammatical rules have been formulated, a fluent speaker almost never thinks of them.

The situation is rather different as regards learning to play chess. As Ryle says (1949, pp. 40-41):

"Consider a boy learning to play chess. … this generally involves his receiving explicit instruction in the rules. He probably gets them by heart and is then ready to cite them on demand. During his first few games he probably has to go over the rules aloud or in his head, and to ask now and then how they should be applied to this or that particular situation."

However, this applies only when the boy is a novice learning to play chess. Once he knows how to play chess, it no longer holds. As Ryle says (1949, p. 41):

---

[1] For further details about this, see … .

 doi:10.20944/preprints202410.1221.v1

"But very soon he comes to observe the rules without thinking of them. … At this stage he might even have lost his former ability to cite the rules. … So long as he can observe the rules, we do not care if he cannot also formulate them."

This then is Ryle's distinction between knowledge how and knowledge that. The distinction we will use is not entirely the same. We want to distinguish knowledge how, not from the whole of knowledge that, but from a subset of knowledge that, which we will call 'knowledge with a rationale'. There are two reasons for this.

First of all, knowledge how usually has associated with it some knowledge that. For example, if Francesca, aged 10, knows how to recognise Uncle John, then she will also know that the man sitting in the armchair in the living room is Uncle John. Suppose further that Francesca, who is a very intelligent child, knows how to speak Italian, play the piano and play chess. Following Ryle's analysis, she is likely to know some of the rules associated with these activities. For example, she will surely know that bishops in chess move diagonally.

Secondly, we come to the question of justification. It is usually easy to check whether someone's claim to know how to do something is justified. All we have to do is ask them to do the thing in question and see whether they carry out the task successfully. For example, if there were some doubt as to whether Francesca knew how to do the things just listed, we could proceed as follows. Regarding the claim that she can speak Italian, we have only to engage her in a conversation in that language. If Francesca converses fluently and grammatically in Italian, making many interesting points, we have justified the claim that she knows how to speak Italian. Regarding the claim that she can play the piano, we have only to ask her to play a piece for us. If she plays a difficult piece with feeling and without making a mistake, we have justified the claim that she knows how to play the piano. Regarding the claim that she can play chess, we have only to challenge her to a game. If she trounces us, we can be sure that she knows how to play chess.

While this is usually the case, there are of course some exceptions. People who know how to do something may on occasion fail because they are suffering from flu, are stressed out by personal problems, or for some similar reason. For example[2], Simone Biles in general knows how to do complicated vaults, but at the 2020 Olympics, she was unable to perform because she got the 'twisties'. If someone claims to possess knowledge how but fails to exhibit it, we must reject their claim. After all, even if that person did have knowledge how in the past, they may have lost it. If, however, like Simone Biles, they start performing well again, we can say that they have regained their knowledge how.

Even taking into account the question of occasional failures, the criteria for establishing knowledge how are relatively simple. The situation is not so simple in some cases of knowing that. For example, in 1750, nearly all expert mathematical astronomers would have claimed that they knew that Newton's mechanics and his law of gravity were correct. However, if we had asked them to justify this knowledge claim, the response would have been quite complicated. They would have offered a rationale involving empirical observations, experiments, and mathematical deductions and calculations. Ryle is aware of this point for he writes (1949, p. 28):

"though it is proper to ask for the grounds or reasons for someone's acceptance of a proposition, this question cannot be asked of someone's skill at cards"

Our next task is to consider the nature of knowledge with a rationale.

3. Knowledge with a Rationale

The concept of knowledge with a rationale is to be found in many philosophers. It will be convenient to begin with the views of Leibniz on this subject. After the publication in 1700 of a French translation of Locke's *An Essay Concerning Human Understanding*, Leibniz set himself the task of writing a rebuttal entitled: *New Essays on the Human Understanding*. However, this interesting work was only published in 1765, nearly fifty years after Leibniz' death. Leibniz writes (c. 1700, p. 143):

---

[2] This exsample was suggested to me by Emily Sullivan.

 doi:10.20944/preprints202410.1221.v1

"From this arises another question, whether all truths depend on experience, that is to say on induction and on instances, or whether there are some which have another basis also."

Of course, Leibniz, being a rationalist, will argue that not all truths depend just on induction and on instances, but some depend on reason as well. Now, it is clear that his arguments are relevant to the knowledge produced by deep neural networks, since, as was remarked above, these work by a form of computer induction. Thus, neural networks embody an empiricist approach and Leibniz's criticisms of that approach therefore apply to the valuation of the output of these networks. Leibniz argues against exclusive reliance on induction as follows (c. 1700, p. 144):

"Now all the instances which confirm a general truth, however numerous they may be, are not sufficient to establish the universal necessity of this same truth, for it does not follow that what happened before will happen in the same way again. For example, the Greeks and the Romans, and all the other peoples of the earth known to the ancients, always observed that before the passage of twenty-four hours day changes to night and night to day. But they would have been wrong if they had believed that the same rule hold good everywhere, for since that time the contrary has been experienced during a visit to Nova Zembla."

Leibniz then goes on to argue that animals are purely empiricist, but that men are superior because they have reason. As he says (c. 1700, p. 145):

"It is in this also that the knowledge of men differs from that of the brutes: the latter are purely empirical, and guide themselves solely by particular instances; for, as far as we can judge, they never go so far as to form necessary propositions; whereas men are capable of demonstrative sciences. This also is why the faculty the brutes have of making *sequences* of ideas is something inferior to the reason which is in man. The sequences of the brutes are just like those of the simple empiricists who claim that what has happened sometimes will happen again in a case where what strikes them is similar, without being capable of determining whether the same reasons hold good. It is because of this that it is so easy for men to catch animals, and so easy for pure empiricists to make mistakes. And people whom age and experience has rendered skilful are not exempt from this when they rely too much on their past experience, as some have done in civil and military affairs; they do not pay sufficient attention to the fact that the world changes, and that men become more skilful by discovering countless new contrivances, whereas the stags and hares of to-day are no more cunning than those of yesterday."

Leibniz develops these arguments in *The Monadology* of 1714, where he writes (pp. 7-8):

"26. Memory provides souls with a kind of *consecutiveness*, which copies reason but must be distinguished from it. What I mean is this: we often see that animals, when they have a perception of something which strikes them, and of which they had a similar perception previously, are led, by the representation of their memory, to expect what was united with this perception before, and are carried away by feelings similar to those they had before. For example, when dogs are shown a stick, they remember the pain which it has caused them in the past, and howl or run away.

…

28. Men act like brutes in so far as the sequences of their perceptions arise through the principle of memory only, like those empirical physicians who have mere practice without theory. We are all merely empiricists as regards three-fourths of our actions. For example, when we expect it to be day tomorrow, we are behaving as empiricists, because until now it has always happened thus. The astronomer alone knows this by reason."

Leibniz here clearly demarcates knowledge which is based purely on induction from observations from the kind of knowledge which the astronomer possesses and which is based on reason. It is this second kind of knowledge which we are describing as 'knowledge with a rationale'. However, we will not pursue Leibniz' analysis of this kind of knowledge in detail, but rather skip some centuries and consider a passage of Popper's which uses a very similar example to that of Leibniz. Popper says (1972, p. 20):

"Thus I assert that with the corroboration of Newton's theory, and the description of the earth as a rotating planet, the degree of corroboration of the statement s 'The sun rises in Rome once in every twenty-four hours' has greatly increased. For, on its own, s is not very well testable; but

 doi:10.20944/preprints202410.1221.v1

Newton's theory, and the theory of the rotation of the earth are well testable. And if these are true, *s* will be true also."

I prefer to use the term 'confirmation', meaning empirical confirmation by observations and experiments, rather than Popper's term 'corroboration'. However, with this alteration, Popper's general point can be put as follows. Let us start with Leibniz' original example of a possible law (the law of the sun's setting and rising), namely: 'Before the passage of twenty-four hours day changes to night and night to day.' This law was obtained by induction from countless observations in the ancient world. These observations confirmed the law, and we shall call this type of confirmation: *direct confirmation*. Now the law of the sun's setting and rising can with some significant qualifications be deduced from Newton's theory, which I take to include the theory of rotation of the earth. The significant qualifications, which are alluded to by Leibniz, are that the law does not apply to some regions of the earth near the poles at some times of the year. If we modify the law to take account of these qualifications, then it obtains, in addition to its direct confirmation, some *indirect confirmation*. Newton's theory is confirmed by observations on the planets, on the tides, on the motion of pendula, on the motion of projectiles etc. Since these observations confirm Newton's theory and since the modified law of the sun's setting and rising is derivable from Newton's theory, it follows that this modified law is indirectly confirmed by these observations on the tides, on pendula, etc. which, prior to the introduction of Newton's theory might well have seemed completely irrelevant to it.

This example is typical of what we are calling 'knowledge with a rationale' in the natural sciences. Knowledge obtained purely by induction from observations is not in itself knowledge with a rationale, but it can be turned into knowledge with a rationale by explaining it by some theory. Usually, the explanatory theory corrects the original empirical law in some ways, and it also provides some indirect evidence for the modified empirical law. This notion of 'rationale' is rather different from that of Leibniz, but they are not unconnected. Leibniz, as we have seen, speaks of 'demonstrative sciences' which men but not animals are capable of producing. We can also agree with Leibniz that knowledge with a rationale is superior to knowledge based solely on induction from observations. Knowledge with a rationale is based on theoretical explanations which both indicate the limitations of empirical generalisations and also provide indirect as well as direct confirmation for modified versions of these empirical generalisatons. Despite these seemingly convincing arguments, knowledge with a rationale has been challenged recently by Chris Anderson in his famous and provocative 2008 article: 'The End of Theory: The Data Deluge Makes the Scientific Method Obsolete'.

Anderson writes:

"The new availability of huge amounts of data along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all."

However, if explanatory theories are no longer produced, then knowledge with a rationale in science will also disappear. But do huge amounts of data really make knowledge with a rationale unnecessary? Leibniz' example of the 'law' that before the passage of twenty-four hours day changes to night and night to day shows that this is not the case. A massive amount of data could in principle have been collected from all the civilised peoples on earth for thousands of years. All of which would have confirmed this law, since those regions of the earth where it breaks down (such as Nova Zembla) did not have any civilised inhabitants, or, for the most part, any inhabitants at all. Yet the law was not correct, and we can see that it was worth developing astronomical theories to explain this law, because they showed its limitations.

Similar considerations apply to Anderson's more specific claim that "correlation supersedes causation". Anderson also says, quite correctly:

"Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y … . Instead, you must understand the underlying mechanisms connecting the two."

 doi:10.20944/preprints202410.1221.v1

This is indeed part of scientists' training and, contrary to what Anderson claims, a valuable part, as we can show by a simple example. Heavy drinking is strongly correlated with lung cancer, but most researchers would not regard this correlation as causal in character. It almost certainly arises because heavy drinking is strongly correlated with heavy smoking and there is a causal connection between heavy smoking and lung cancer. Recognising which correlations are causal in character and which are not, is important in practice. Giving up heavy smoking for someone who continues to drink heavily will indeed reduce the risk of lung cancer, but giving up heavy drinking for someone who continues to smoke heavily will not reduce the risk of lung cancer to any significant extent.

It is subtle matter to decide whether a particular correlation is causal or not. Theoretical considerations and indirect evidence need to be brought in. So, genuine causal knowledge is always knowledge with a rationale. Anderson's claim is that, with huge amounts of data, causal knowledge becomes unnecessary, and that correlation is all that is required. However, it is easy to give counterexamples to this thesis. The standard example of a correlation which is not causal is the following. The reading of a barometer falling in value is strongly correlated with rain occurring. However, the connection is not causal. If someone reduces the value of the reading by fiddling with the barometer, this will not produce rain. Now this example can easily be generalised to fit our era of big data. Suppose all smart phones are equipped with sensors which can detect rain, and with an app, which, in conjunction with the sensors, can act as a barometer. Suppose that the data from the app and the rain-detecting sensors is harvested from millions and millions of smart phones by some large corporation. A correlation between the reading of the barometer app falling and rain occurring is established on the basis of huge amounts of data, but this does not make the correlation causal. If a hacker breaks into the system and makes the value of the barometer app fall, this will not produce rain. So, very large amounts of data do not make knowledge with a rationale redundant in the sciences.

The situation regarding the relation between correlation and causation is very similar to that, analysed earlier, between the law of the sun's setting and rising and Newton's theory. This can be illustrated by the following example.[3] To investigate the relationship between diet and coronary heart disease, Ancel Keys carried out an extensive study in seven countries. The study was begun between 1958 and 1964 and in 1970, Keys published the five year results. One finding concerned the relationship between x (the percentage of the diet calories which came from saturated fat) and y (the percentage of the cohort eating that diet who had died of coronoary heart disease). The result showed a close approximation to a linear regression model with a correlation of 0.84. Keys thought that this provided strong evidence of the the claim that a diet high in saturated fat causes coronary heart disease. However, he was well aware that *correlation is not causation*, and sought additional evidence to establish the causal claim.

One of the ways in which a correlation can fail to be a causation is because of the existence of a 'confounder'. This is illustrated by our earlier example of the correlation between heavy drinking and lung cancer where the confounder was heavy smoking. Could there be a confounder in Keys' diet study? The Japanese cohort had a diet which contained much less saturated fat than the American cohort, and the death rate from coronary heart disease was also much lower among the Japanese than among the Americans. Here, however, there could have been a genetic confounder. The Japanese might have a set of genes which protected them against coronary heart disease and which the Americans lacked. However, there was an easy way to test the hypothesis of a genetic confounder. Keys had only to compare Japanese living in Japan and eating a traditional Japanese diet with Japanese who had emigrated to the U.S.A. and started eating a typical American diet. It turned out that the latter category of Japanese had rates of coronary heart disease more or less the same as those of other Americans, and this ruled out a genetic confounder.

Another way of establishing causality is to try to undertand the mechanisms (if any) connecting the correlated variables. In the present case this was done very successfully through experiments on rabbits fed an unusual diet and also from observations obtained from autopsies of humans. These

[3] For more details about this example, see… .

 doi:10.20944/preprints202410.1221.v1

led to the elucidation of the following mechniam. A diet high in saturated fat produces a raised level of LDL cholesterol in the blood, which leads to infiltration of some parts of the arterial wall by lipids, which are then absorbed by macrophages, and become foam cells. This brings about the full development of atherosclerotic plaques. This mechanism explains the occurrence of coronary heart disease which is produced by atherosclerotic plaques forming in the coronary arteries. The mechanism also leads to a generalisation of the proposed causal law, since a diet high in saturated fat can cause any form of atherosclerosis, and not just coronary heart disease. Strokes, for example, are another form of atherosclerosis.

A third way of checking causal claims is the use of randomized controlled trials, since randomization provides a technique for dealing with unknown confounders. A double blind randomized controlled trial was carried out in the USA in the period 1959-1968, and sure enough it showed that the group which ate the diet low in saturated fat had a statistically significant decrease in the number of atherosclerotic events.

This example is very similar to that of the law relating to the sun's setting and rising and Newton's theory considered earlier. The starting point in the saturated fat case is an observed correlation. This is explained by postulating a causal link between the correlated variables, and an investigation proceeds as to whether this causal link is correct. In the course of this investigation, the original causal law is changed (generalised) and it is then confirmed by the results of a variety of experiments and observations, quite different from those which confirmed the original correlation. This new evidence included the testing of possible confounders, the establishment of physiological mechanisms, and the result of randomized controlled trials. At the end of the investigation, a rationale has been provided for a causal law which explains the observed correlation.

Knowledge with a rationale is also important in the law. It would obviously be unjust to send someone to prison, unless we can validly claim to know that he or she is guilty. This knowledge must be capable of being justified by producing an explicit rationale for the guilty verdict, and so should be knowledge with a rationale.

That concludes our account of the distinction between knowledge how and knowledge with a rationale. We will now briefly compare it to Sosa's quite similar distinction.

Sosa writes (1985, pp. 241-2):

"From this standpoint we may distinguish between two general varieties of knowledge as follows:

One has *animal knowledge* about one's environment, one's past, and one's own experience if one's judgements and beliefs about these are direct responses to their impact – e.g. through perception or memory – with little or no benefit of reflection or understanding.

One has *reflective knowledge* if one's judgment or belief manifests not only such direct response to the fact known but also understanding of its place in a wider whole that includes one's belief and knowledge of it and how these come about."

The passages from Leibniz quoted above cast some light on Sosa's expression 'animal knowledge' since Leibniz thought that animals could only gain knowledge by induction from instances. However, the term does not seem to us quite appropriate for our investigation, since we want to consider learning such things as language, or how to play the piano or to play chess. These are all things which, apart possibly from the language of bees, are characteristic of humans rather than animals. This is why we prefer to use Ryle's concept of knowledge how which applies to such examples. Sosa's concept of reflective knowledge, however, seems quite close to our concept of knowledge with a rationale.

Having explained the two kinds of knowledge we want to consider, we must see how they apply to deep learning. Before doing so, however, we must draw attention in the next section to a relevant feature of deep neural networks (and indeed of nearly all neural networks), namely that the results they produce are opaque, in the sense of being unintelligible to human beings.

4. The Opaqueness of Neural Networks

 doi:10.20944/preprints202410.1221.v1

One of the most striking features of deep learning networks is that they are very hard, if not impossible, for human beings to understand. This is not because the mathematics involved is very complex. The mathematics of an individual artificial neuron is actually quite simple. The inputs are combined linearly with weights, and a so-called activation function, which is non-linear is then is then applied. The activiation functions used are very simple and would be familiar to any student of high school mathematics. The power of deep learning does not come from the complexity of the mathematics, but from the size of the models: the large number of neurons involved. The action of these many neurons can be far beyond what we humans are capable of understanding. Computers are very good at handling huge numbers of simple computational units, but humans, with our limited working memories, can find them completely opaque.

Another way of putting the point is to say that deep learning networks essentially learn a function which maps the inputs to the output. This function is expressed by the final state of the network with its various weights, but this state is not comprehensible to human beings and cannot be reduced to a simple formula which is comprehensible. This situation can be illustrated by a well-known example, namely the network which, in 2012, started the deep learning revolution.

An annual competition for computer vision researchers: the ImageNet prize, had been set up in 2009, after the publication of ImageNet which had been had been developed by Feifei Li at Stanford University. The competition was called the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC). ILSVRC used a subset of ImageNet with roughly 1000 images in each of 1,000 categories. In all, there were roughly 1.2 million training images, 50,000 validation images, and 150,000 testing images. In the 2012 competition, the power of deep learning became apparent to the community when two of Geoffrey Hinton's students Alex Krizhevsky and Ilya Sutskever (Krizhevsky, Sutskever, & Hinton, 2012) used a deep convolutional neural network which not only performed best but beat the best current network by over 10 percentage points, a massive improvement. It was obvious that a new technology much more powerful than its predecessors had arrived.

The network which Hinton and his students used to enter the 2012 ImageNet Competition consisted of eight learned layers – five convolutional and three fully connected. It had 650,000 neurons and 60 million weights. The network took five to six days to train using the training set of 1.2 million images. It is surely obvious that a system consisting of 650,000 artificial neurons arranged in eight learned layers and with 60 million parameters or weights is not comprehensible to human beings and must be considered as opaque. Actually much smaller neural networks are still incomprehensible while, more recently, much bigger neural networks have been created.

Attempts have been made to overcome the opacity of deep neural networks by what is known as 'Explainable AI'. Although it is impossible to predict the future of any research, the explainable AI research programme has not so far been very successful. Consequently, it is correct to consider most neural networks in use today as being opaque and unintelligible to human beings. In the next two sections, we will consider the consequences of this for the knowledge produced by such networks.

5. Deep Learning Does Provide Knowledge How

In this section, I will make the positive claim that neural networks do provide knowledge how. François Chollet in his 2018 book on deep learning gives on pp. 11-12, a list of eleven breakthroughs achieved by deep learning, all in areas which had proved historically difficult for machine learning. The following are eight of his eleven examples.

1. Near-human level image classification
2. Near-human-level speech recognition
3. Near-human-level handwriting transcription
4. Improved machine translation
5. Improved text-to-speech conversion
6. Near-human-level autonomous driving
7. Ability to answer natural-language questions
8. Superhuman Go playing

 doi:10.20944/preprints202410.1221.v1

These eight examples have been chosen because what the machine does can be compared directly with what humans do. For example, in 4, machine translation can be compared with human translation, and similarly with the other examples. Now the human skills which deep learning is emulating are all obviously examples of knowledge how. This confirms the idea that neural networks can produce knowledge how. Moreover, this offers a justification for the use of neural networks in many cases. After all, as we have seen, humans often learn how to do things, such as recognising images, without having the least idea of how they do it. In such cases, it is reasonable to suppose that the learning process has produced changes in the human brain which enable the person to carry out the activity in question. If that is all right as far as humans are concerned, then surely it is all right to use neural networks if they too can be trained to perform the activity in question. Indeed, though neural networks are indeed opaque, we probably know more about how they work than about how the human brain works.

There is, however, one counterargument which may be important in some cases. Although neural networks very often perform statistically better than humans, they do, like humans, make mistakes in a small percentage of instances; but the mistakes made by neural networks are very different from those made by humans. Some examples are given in Metz 2021, p. 212. It is possible to change a few pixels in a photograph of an elephant, a change which is imperceptible to the human eye, but which leads a trained neural network to reclassify the image as a car. Another curious example is that by putting a few Post-it notes on a stop sign, you can render it invisible to the neural network of a self-driving car. The fact that the errors of neural networks are different from those of humans is not really so surprising. After all, neural networks are very different from the human brain and are trained in a different way. Now in some applications, a small percentage of curious errors is not a problem. For example, in translation, the final version will undoubtedly be monitored by a human, and, if the machine has made a bizarre mistake, this will be easy to spot and so to correct by a human. In other applications, however, the possibility of strange errors is much more serious. Although humans do often crash their cars, the mistakes they make, and the kind of resulting crash are quite familiar. However, self-driving cars can crash for quite unaccountable reasons and in a way which causes major disruptions. Thus, self-driving cars are unlikely to become a reality. Besides, why should we want self-driving cars anyway? Contemporary thinking is mainly in favour of reducing the number of cars and laying the emphasis on public transport, walking and cycling.

This defence of deep neural networks or DNNs is similar to that given by Buckner in his 2023. Buckner considers some criticisms of DNNs, one of which is that (2023, p. 688) "Deep neural networks are not interpretable". As he goes on to say:

"Another common lament holds that DNNs are 'black boxes' that are not 'interpretable' … or not 'sufficiently transparent'."

Buckner thinks that this and other criticisms of DNNs arise because of what he calls 'anthropofabulation' by which he means roughly that there is a tendency to believe in fantasies about the nature of human cognition. Once we form a more realistic view of human cognition, it becomes clear that it is no better than what DNNs can achieve. As he puts it (2023, p. 693):

"Once the anthropofabulation in these critiques is exposed, they no longer clearly support the conclusion that deep learning systems and human brains are performing fundamentally different kinds of processing – and indeed might teach us hard lessons about our own cognition as well."

When rebutting the criticism about the opaqueness of DNNs, he argues (2023, p. 698) that "Human decision making is also opaque".

Up to a point, I would agree with this. In instances of knowledge how, human beings may perform tasks with great skill, and yet have not the slightest idea of how they do it. The successful performance may indeed involve decisions whose origin is opaque. However, not all human decision making is of this character. Consider a judge deciding on a legal case. He or she has not only to give a verdict, but also to provide a rationale for that verdict. Far from being opaque, this rationale has to be clear and well-argued as it may be challenged by an appeal court. This naturally brings us to the question of knowledge with a rationale which will be considered in the next section.

6. But Can Deep Learning Provide Knowledge with a Rationale?

 doi:10.20944/preprints202410.1221.v1

It seems very doubtful whether deep learning can provide knowledge with a rationale. Let us consider the scientific examples given in section **3**. Empirical generalisations such as the law of the sun's setting and rising were indeed obtained by induction from observations. However, providing a rationale for them, involved explaining (and at the same time modifying them) using a higher level theory, such as Newton's theory. This theory enabled the laws to obtain indirect as well as direct empirical confirmation. Much the same applied in the case of observed correlations which are given a rationale by being explained in terms of causal connections. Establishing these causal connections required bringing in empirical evidence of different kinds and so increased the empirical confirmation of the original correlation. Now all these procedures for providing a rationale, involved placing the original result obtained by induction in the main body of human knowledge and explaining it by higher level theories, causal laws etc. If the original result is opaque and unintelligible to humans, it would seem impossible to carry out these procedures for providing a rationale. Similar considerations apply in the legal case, and so the general conclusion would seem to to be that deep learning cannot provide knowledge with a rationale.

This conclusion must, however, be qualified by some points which Sullivan makes in her 2022. In this paper, she considers the interesting example of the neural network model 'Deep Patient' to be found in Miotto *et al* (2016). Sullivan describes it as follows (2022, pp. 109-10):

"This model takes as inputs large amounts of patient medical data and gives as an output a generalizable patient representation that can be used to predict future medical problems. The model provides surprising results. It is able to predict a wide array of medical problems, such as schizophrenia, attention-deficit disorder, and severe diabetes, with a higher degree of accuracy than competing predictive models. However, with this increased accuracy comes increased opacity."

Sullivan goes on to make a significant and novel point that the problem here is not just opacity but also what she calls 'link uncertainty'. Speaking of Deep Patient and other DNN (deep neural net) models, she says (2022, p. 110):

"It is not simply the complexity or opaqueness of DNN models that limits how much understanding they provide. Instead it is the level of 'link uncertainty' present – that is, the extent to which the model fails to be empirically supported and adequately linked to the target phenomenona – that prohibits understanding."

What Sullivan describes as 'link uncertainty' corresponds to low empirical confirmation, since if a model has low empirical confirmation, it may not be linked adequately to the target phenomena which it is supposed to describe. So, to reduce link uncertainty, we have to increase the empirical confirmation of a model. In the case of human science with humanly comprehensible models, our earlier analysis showed a number of ways in which this could be done. A model could be explained by a higher level theory which is empirically confirmed by a variety of different types of evidence, and then this empirical confirmation carries over to the model. If the model is a statistical correlation, it could be explained by a causal law which is again empirically confirmed by a variety of different types of evidence. However, these ways of increasing the empirical confirmation of a model cannot be applied in the case of a DNN model which is unintelligible to human beings.

This seems to block the possibility of reducing link uncertainty for DNN models. However, Sullivan shows, using her example of Deep Patient, that there is a way of increasing the empirical confirmation of such models. This is based on the idea that, although the DNN model may be opaque, the data set used to train and validate the model is perfectly intelligible to human beings. Indeed this data set was prepared by humans. The data in it would be interpreted in terms of humanly comprehensible theories which involve background knowledge. This knowledge contained in the data can, so to speak, be carried over to provided empirical confirmation or disconfirmation of the DNN model which has been trained on the data. As Sullivan says (2022, p. 122):

"Once the model is trained, the modeller still has a general idea of how the finalized model works in virtue of having knowledge about how the model was trained and validated."

 doi:10.20944/preprints202410.1221.v1

She goes on to illustrate this with her Deep Patient example. This was trained on patient medical data, but nonetheless failed to give satisfactory results in some cases. Sullivan shows how this can be explained by a consideration of the training data. As she says (2022, p. 125):

"the modellers found that the model had trouble predicting certain diseases that otherwise should have been predicted with ease, such as diabetes mellitus without complications. Their hypothesis was that since the screening process of diabetes often occurs during routine check-ups, the frequency of those tests was not a valid discriminating factor. This suggests that the model in part tracks proxies of disease development, such as previous physicians' decisions to carry out a diagnostic test. Given this, there is still link uncertainty that prevents understanding of real world instances of disease development."

This suggests that the model could be improved by limiting the data to those which report the occurrence of what are known to be causal factors for the disease. The knowledge thus embodied in the data would then carry over to the model, reducing its link uncertainty. The creators of Deep Patient seem to have used this principle, for, as Sullivan says (2022, pp. 124-5):

"In the deep patient case, the model is greatly influenced by existing empirical evidence concerning diseases. The modellers made particular determinations of which medical problems to include in their predictions and which ones to exclude. For example, they did not seek predictions of HIV because of the large behavioural aspect to the disease (Miotto *et al* [2016], p. 5). Having prior knowledge about which records are salient for medical diagnosis helped lead to the success of the model."

Sullivan is undoubtedly correct that the careful use of background knowledge in preparing the data on which a DNN model is trained can help towards a partially understanding of the model and an increase in its empirical confirmation. Yet, in most cases, this seems to me to fall short of providing the kind of rationale for DNN models which can be obtained in the case of humanly comprehensible models.

Thus, the conclusion of this section is that, on the whole, deep learning does not provide knowledge with a rationale, but that two qualifications should be made to this claim. First of all, although the explainable AI research programme has not so far achieved very striking results, it may do so in the future. Secondly, as Sullivan points out, some understanding of DNN models and some empirical confirmation of them can be obtained by a consideration of the data on which they were trained.

7. Conclusions

Our general conclusion, then, is that deep learning does provide knowledge how, but, in most cases and with the qualifications just mentioned, it does not provide knowledge with a rationale. Now, there is a great deal of concern at the moment that deep learning may be applied in cases for which it is not suitable. The distinction between knowledge how and knowledge with a rationale may help to shed light on this problem. Let us consider the example of selecting someone for a job. It is perfectly possible to train a deep learning system to carry out this selection. We need only feed in a training set consisting of a large number of earlier job selections, where in each case the applications of all the candidates who applied are given and also information on who was the successful candidate. Deep learning would then be able to discover characteristics of successful applicants from this data, and so select candidates in future cases. Needless to say, most people would not approve of this particular application of contemporary AI!

One problem is that aggrieved candidates who think that they have been discriminated against unfairly can, and sometimes do, bring legal actions against those who carried out the selection process. These selectors have therefore to provide a rationale for their decision, and this is just what a deep learning system cannot do. Such a deep learning system might well have learnt biases based on race and gender which were acceptable in the past but are no longer so. Without an explicit, humanly comprehensible, rationale for a decision, the existence of biases learnt form the past may not be detected.

 doi:10.20944/preprints202410.1221.v1

References

1. Anderson, C. (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired Magazine*.
2. Buckner, C. (2023) Black Boxes or Unflattering Mirrors? Comparative Bias in the Science of Machine Behaviour, *The British Journal for the Philosphy of Science*, **74**(3), pp. 681-712.
3. Chollet, F. (2018) *Deep Learning with Python*, Manning.
4. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012) ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *2*.
5. Leibniz, G.W. (c. 1700) *New Essays on the Human Understanding*. English translation by Mary Morris of selections in *Leibniz: Philosophical Writings*, Dent, 1961, pp. 141-191.
6. Leibniz, G.W. (1714) *The Monadology*. English translation by Mary Morris in *Leibniz: Philosophical Writings*, Dent, 1961, pp. 3-20.
7. Metz, C. (2021) Genius Makers. The Mavericks Who Brought AI to Google, Facebook and the World, Penguin Edition, 2022.
8. Miotto, R., Li, L., Kidd, B. and Dudley, J. (2016) Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records, *Scientific Reports*, **6**, pp. 1–10.
9. Otsuka, J. (2023) Thinking About Statistics. The Philosophical Foundations, Routledge.
10. Popper, K.R. (1972) *Objective Knowledge. An Evolutionary Approach*, Oxford University Press.
11. Ryle, G. (1945-6) Knowing How and Knowing That, *Proceedings of the Aristotelian Society*, **46**, pp. 1-16.
12. Ryle, G. (1949) *The Concept of Mind*, Hutchinson, 1960.
13. Sullivan, E. (2022) Understanding from Machine Learning Models, *the British Journal for the Philosphy of Science*, **73**(1), pp. 109-133.
14. Sejnowski, T.J. (2018) *The Deep Learning Revolution*, The MIT Press.
15. Sosa, E. (1985) Knowledge and Intellectual Virtue, *Monist*, **68**(2), pp. 226-245.