PRINCIPIA **26**(3): 597–614 (2022) doi: 10.5007/1808-1711.2022.e84419
Published by NEL — Epistemology and Logic Research Group, Federal University of Santa Catarina (UFSC), Brazil.

BIG DATA AND DEEP LEARNING MODELS

DANIEL SANDER HOFFMANN
Universidade Estadual do Rio Grande do Sul, BRAZIL
daniel-hoffmann@uergs.edu.br

Abstract. Although deep learning has historically deep roots, with regard to the vast area of artificial intelligence and, more specifically, to the study of machine learning and artificial neural networks, it is only recently that this line of investigation has developed fruits with great commercial value, starting to have thus a significant impact on society. It is precisely because of the wide applicability of this technology nowadays that we must be alert, in order to be able to foresee the negative implications of its indiscriminate uses. Of fundamental importance, in this context, are the risks associated with collecting large amounts of data for training neural networks (and for other purposes too), the dilemma of the strong opacity of these systems, and issues related to the misuse of already trained neural networks, as exemplified by the recent proliferation of deepfakes. This text introduces and discusses these issues with a pedagogical bias, thus aiming to make the topic accessible to new researchers interested in this area of application of scientific models.

Keywords: artificial intelligence • artificial neural networks • big data • black boxes • deepfakes • deep learning

RECEIVED: 15/10/2021 REVISED: 30/05/2022 ACCEPTED: 26/07/2022

1. Introduction

Deep learning (DL) entails creating, training, and using deep neural networks, which are nothing but good old-fashioned artificial neural networks (ANNs) that happen to contain a relatively high number of *hidden* (i.e., internal) layers of artificial "neurons." Hence, the origins of deep learning can ultimately be traced back, alongside the whole field of artificial neural networks, to the pioneering work of McCulloch and Pitts, in the first half of the 1940s. Therefore, it is safe to say that deep learning is the latter manifestation of this area of research, bringing with it new tools, concepts, and techniques.

It is known that the relative importance and visibility of artificial neural networks, in the academy and in industry, has oscillated significantly along the decades, in response to the so-called "AI winters" (there seems to be some disagreement about this topic between historians of science, however). Indeed, it is frequently mentioned, in the artificial intelligence (AI) and machine learning (ML) technical literature, that

the subject of artificial neural networks has repeatedly found itself in- and out-of-fashion, in the past. Part of the blame was due to the poor performance of computers back then, that limited the size of the artificial neural networks that could be constructed and trained, hence their applicability. And part of the blame can certainly be credited to the lack of efficient training algorithms.

This situation has changed radically in recent years, because of the availability of computers that are far more powerful than their predecessors, as well as due to the use of specialized and fast hardware that go by the name of FPGAs, GPUs, TPUs and so on. With more computer power available, research in artificial neural networks gained impetus and, following an impressive result obtained by deep neural networks in 2011, there have been still more remarkable advances in the area, many of them immediately turned into highly profitable commercial applications, that are right now percolating our society at a fast pace. So far so good but, as we all know, there is no (such thing as a) free lunch, so it is important that we properly understand the negative aspects of this technology and prepare ourselves in order to be able to anticipate still unseen potential hazards.

This text introduces the reader to some of the positive aspects surrounding the creation, training and application of deep neural networks, while concomitantly addressing their *shortcomings*, namely: (i) that huge deep learning models share some of the issues that plague *big data science* in general, (ii) that these models are revealing themselves as appallingly efficient tools for "deepfaking" and similar malicious practices, and (iii) that it is frequently hard to explain the inner workings of huge trained deep learning models (the black box problem), something that may eventually have consequences for how science and technology are practiced.

The paper is not to be taken as an updated primer or a comprehensive review about artificial intelligence, neural network models (deep or otherwise), deepfakes, opacity in machine learning models, or big data. The goal instead is to offer an overview of these topics to new researchers interested in this important area of model application. Still regarding modeling, here and there, in the text, I draw attention to some of the main modeling approaches or interpretations that one usually finds in the literature, when studying deep learning and other machine learning techniques. Finally, the reader should also not expect to find here much of a dialectics between philosophers working in this area of investigation. Indeed, I consciously take a different approach, aimed at highlighting technological aspects of machine learning (mainly deep learning) that bear relation to models, while concomitantly addressing shortcomings that may be of interest to philosophers that are new to this subject matter.

2. Deep Learning

In what follows, I first review some commonly accepted definitions of artificial intelligence, machine learning, and deep learning, then I briefly mention a handful of historical facts, and finally I address some topics that I consider to be of interest to new investigators.

The area widely known as artificial intelligence, that aims to understand and build intelligent agents, is among the fastest-growing technological fields, and generates over a trillion dollars in revenue each and every year (Russell & Norwig 2020). A way of defining artificial intelligence is to say that it is a cross-disciplinary approach to create, model, replicate and understand intelligence (and cognitive processes) of different forms, using mathematical, computational, logical, mechanical, and biological devices and principles. It can be considered as a branch of cognitive science as well, because it frequently develops explanatory models of animal and human cognition (Frankish & Ramsey 2014), as we are going to see below.

A given agent is said to *learn* if observations about the world improve its ensuing performance. If the agent in question is a machine, as opposed to a person, then one is dealing with machine learning. Hence, machine learning is a subfield of artificial intelligence that investigates how machine performance can be improved by means of experience. During the learning process, *the machine observes the data, builds a model based on this data, and uses the model as a hypothesis about the world, as well as a way to solve problems*. This applies both to an autonomous agent like a mobile robot navigating a complex environment, and to a static desktop computer predicting stock market prices (Russell & Norwig 2020). *Data modeling* is thus the first use of "models" in artificial intelligence that I am going to highlight in this paper.

So much for machine learning in general. Deep learning, by its turn, is a term reserved for all those machine learning implementations that make use of *many layers* (hence the word "deep") of simple and adjustable computing elements. Another way to put it is to say that deep learning is a set of machine learning techniques in which hypotheses materialize, so to speak, as intricate algebraic structures with adjustable connection strengths (Russell & Norwig 2020).

Given that most deep learning systems are, in practice, nothing but many-layered neural networks, it follows that their origin can, in a sense, be traced back to the birth of the whole artificial neural networks field, namely to the original work with neural network models developed by the American cybernetician and neurophysiologist Warren Sturgis McCulloch, and the American logician Walter Harry Pitts, Jr., in the early 1940s (McCulloch & Pitts 1943).

Before going on, I would like to offer a short description of neural networks and their main components, alongside some remarks about how they "learn." This is hopefully going to help in understanding what comes later in the text. First of all, ar-

tificial neural networks are loosely inspired by the architecture of the animal nervous system (a "natural" neural network), where structures known as synapses account for the interconnection between specialized cells called neurons (Kandel; Barres; Hudspeth 2013). In an artificial neural network, each interconnection (link) between artificial neuron-like elements (units) has a given numerical value or "weight" (alternatively, synaptic weight) that accounts for their degree of connectivity (strength). The activation of each neuron in a given artificial neural network thus depends on the state of all the other neurons it connects with (Thagard 1990; Haykin 2009). It is worth mentioning again that this type of network is conventionally structured into more than one layer of neurons: there is usually the input (first) layer, the output (last) layer and one or more "hidden" (intermediary) ones. I also note that structures like neural networks are usually graphically represented as circles with straight lines between them, and it must be said as well that there are many different possible configurations, each one representing a distinct architecture. I recommend, for the interested reader, the chart produced by van Veen & Leijnen (2016), that is a nice summary of such graphical representations.

In order to learn how to perform a given task, an artificial neural network must necessarily be trained. Learning, here, amounts to fine-tuning, in a proper way, the numerical weights of its interconnections, with the goal of improving its overall performance in the context of a task to be ultimately assigned to it (e.g., to correctly locate and name objects in an image). The textbook learning method applied to artificial neural networks is known as backpropagation, but it is still unclear (Seung & Yuste 2013) if it bears any relation to the learning mechanisms of natural systems such as the monkey brain. In backpropagation (that falls into the category of the so-called "supervised learning"), an automated "teacher" informs, so to speak, the output neurons (again, those pertaining to the last layer) of the network how close they are to yield the correct answer, and small adjustments to the connection weights are subsequently "propagated back" (hence the name of the technique) throughout the whole neural network. This procedure raises the chance that the correct answer will subsequently be obtained (Thagard 1990; Haykin 2009).

The philosopher Paul Thagard (1990) came forward with a simple but interesting example of a small three-layer neural network designed to distinguish between cows, goats, and horses. The network has an input layer containing three different neurons (labeled "mane," "udder," and "tail"), an output layer also containing three different neurons ("horse," "cow," and "goat"), and a single intermediary hidden layer containing only two unlabeled neurons. Each input neuron connects to each neuron of the hidden layer, and each neuron of the hidden layer connects to each neuron of the output layer. There is no connection between neurons that belong to the same layer, and there is no direct connection between neurons of the input and output layers. All the neurons have activation values that account for the presence or absence of

features, being thus active ("on") or inactive ("off"), respectively.

After being presented with the architecture of the neural network, one may rightfully wonder how it works. In fact, this is rather straightforward to understand, due to the simplicity of this particular structure, and it goes as follows (Thagard 1990): animal characteristics (or "features") are presented to the neural network by activating correspondent neurons from the first layer (input neurons). For example, if an animal has mane, then the neuron in the first layer that refers to the presence of mane is triggered (it changes from "off" to "on") and, ultimately, another neuron with an animal label should also trigger in the last layer, thus identifying a particular animal (hopefully one that has a mane, like a horse). If the neural network is still not trained, however, it is going to give completely wrong answers most of the time, so it must, first of all, learn how to output the correct animal when fed with a given animal feature (or more than one).

The network is effectively trained by being shown different samples of animal features, that activate one or more input neurons. The "teacher," by its turn, detects whether the correct output neuron is being activated or not in response to the "stimuli," providing corrections as needed. To use an example chosen by Thagard himself: if the input neurons referring to "tail" and "mane" are both triggered, then the output neuron referring to "horse" is expected to trigger in response. If this does not happen, it is up to the backpropagation algorithm to readjust the weights of the network connections according to a certain well-established procedure (that is beyond the scope of this introductory essay). The weight adjustments occur in the connections between the output neurons and the neurons of the hidden layer, and also between the latter and the input neurons. As a result of a usually long iteration process, the neural network will eventually learn to discriminate the animal features present in the sample, correctly identifying the correspondent animals most of the time.

Going back to the historical aspects of the field, if we want to be more precise about the roots of deep learning, i.e., contemporary neural networks research and development, we must say that it has to do with some specific experiments performed in the 1970s. And in the late 1990s, the French computer scientist Yann LeCun and coworkers managed to produce important results in the area. More specifically, they used the backpropagation algorithm (that we saw before) to train a deep neural network to perform automatic handwritten digit recognition. But it is widely acknowledged that the field really took off in 2011, first in the area of speech recognition, then in visual object recognition. A few months later, in 2012, a deep learning system developed by Geoffrey Hinton and collaborators won the ImageNet competition, classifying images into one of a thousand categories, with a far better performance than its competitors (that relied on other methods). From that year onwards, deep learning systems reportedly outperformed humans in many visual tasks, with great progress shown as well in language translation, medical diagnosis, speech recogni-

tion and game playing (Russell & Norwig 2020).

Readers who are looking for a detailed chronology containing all the landmarks of the field of artificial neural networks may find it in the works of Frankish & Ramsey (2014) and of Goodfellow, Bengio & Courville (2016). Now let us address some important technological facts.

It is important to call attention to the fact that deep neural networks that are under training place a high demand on computational resources, given the need to process highly parallelized vector and matrix operations (Russell & Norwig 2020). It is thus important that the philosopher dealing with this area of inquiry gain at least some firsthand insight into the computational procedures and time scales involved, among other aspects. I also submit that there is no true substitute here for effectively engaging in the process of constructing, training, and applying a few deep learning models. This can even be freely accomplished online in sites such as *Kaggle* or *Google Colaboratory*, inter alia. It is really crucial for the investigator willing to work in the area to acquire the aforementioned insights, including a good understanding of the computational demands related to the training of deep neural networks.

The computational demands here are not to be overlooked: while a standard computer processor can provide up to 10^{10} operations per second, it is fairly common for a deep neural network to consume between 10^{14} and 10^{17} operations per second, hence the processing power usually comes in the form of specialized high performance hardware such as FPGAs (*Field-Programmable Gate Arrays*), GPUs (*Graphics Processing Units*) and, yet more recently, TPUs (*Tensor Processing Units*) (Russell & Norwig 2020; Berggren et al. 2020; Bernard 2021). Note that TPUs are machine learning "accelerators" especially developed by Google around 2015 for the sake of training deep neural networks.

It is well known that deep learning models, and machine learning models by and large, generalize better to new examples (i.e., can deal with examples that were not present in the original training dataset) when trained on more data. But even though there is an abundance of raw data available in many areas, waiting to be analyzed (see the discussion about big data in the next section), the same is not necessarily true regarding properly structured training datasets. One solution to this problem (Goodfellow; Bengio; Courville 2016) is to generate synthetic data, thus growing the size of the training dataset. Dataset augmentation is very useful for training classifiers, especially for solving a particular classification problem, namely object recognition. It is also effective for training speech recognition models.

As Russell & Norwig (2020) point out, the overall architecture of a given neural network is a result of properly choosing its depth (number of hidden layers), width (number of neurons at each hidden layer — the width correlates with how many distinct features the network is dealing with at each level of increasing abstraction), in-layer connections and (at last) between-layers connections. Besides, a production-

ready, well-trained neural network, will necessarily contain appropriate values for all its activations, weights (that are learnable parameters, as we saw before), and biases. Biases, that I did not mention before, are also learnable parameters. To make a biological analogy, a bias would correspond to a threshold above which a biological neuron in the brain of an animal is expected to "fire" (Bernard 2021).

A well-known empirical finding in deep learning can be summarized as saying that, when two deep neural networks with approximately equal numbers of weights are compared, then usually the network with more consecutive hidden (i.e., internal) layers of neurons generalizes better. Apart from this apparent fact, it can be argued that we currently lack a general mathematical theory that relates "form" (network architecture) and "function" (task) of neural networks in a principled way. To put it differently, even if there are some acknowledged pragmatic guidelines for choosing the best network architecture for a given task, choices frequently depend on trial and error, such as when a researcher needs to run many different deep neural networks, only to be able to select the more efficient one (Russell & Norwig 2020).

In machine learning, one expects that a model, trained on a given set, is able to easily generalize to new data, and this is measured by how well the network performs on a test set (a test set, by its turn, must be distinct from the original training set). Hence, some kinds of neural networks are expected to generalize well when applied to images, while others are expected to do the same for audio or text (Russell & Norwig 2020). The computation of the accuracy of a model is an important way of keeping track of how well it performs. Actually, accuracy is a simple metric commonly used for the evaluation of machine learning models that perform classification (that are arguably the most common), and can be stated as being the number of correct answers the model returned, divided by the total number of answers (correct plus incorrect) given by it. In the case of a deep neural network trained as an image classifier, for example, this would amount to counting all the images correctly classified by it, and subsequently dividing this value by all the predictions it made (Charniak 2018; Bernard 2021).

Deep neural networks, sometimes alongside other techniques, have been used in all sorts of applications, and the following list is far from complete: virtual assistants, robotic control and vision, speech recognition, autonomous cars, search engines, complex character behavior in immersive games, automated detection of tumors, unmanned aerial vehicles (UAVs), sentiment analysis, smart advertising, personalized recommendation systems (e.g. for songs and movies), personalized news aggregators, chatbots, prediction of DNA mutations, improved image colorization, fraud detection, prediction of protein folding, financial analysis, bioinformatics, automated theorems proof, fake news detection, music composition, game-playing against human opponents (Go, chess, Jeopardy, etc.), and human-level text writing. At last, but not least, deep learning is being heavily used for analyzing big data coming from

many different areas of scientific inquiry.

Up to this point, we mentioned deep neural networks mostly from the technological perspective, where these *data models* are constructed, trained, and analyzed by engineers and computer scientists, to be subsequently applied as useful tools in many contexts. But according to Cichy & Kaiser (2019), another important trend sees deep neural networks as scientific models for biological cognition, including the study of neural structure and function. As it happens, deep neural networks are very useful for *predicting* and *explaining* cognitive phenomena, e.g., in auditory and visual processing, where cognitive scientists have found that these structures are superior to previous models in predicting brain responses and behavior in humans. Hence, *scientific modeling of cognition* is another use of models, related to artificial intelligence (more specifically to deep learning, in this case), that we must acknowledge.

Cichy & Kaiser (2019) also mention that deep neural networks can contribute to yet another largely disregarded but pervasive (hence important) use of scientific models, namely *exploration*. To begin with, they go on explaining, exploration of models can yield new *hypotheses* due to analogies between deep neural networks and natural neural systems. Besides, deep neural networks can serve as proof-of-principle demonstrations. Finally, these networks may help check the suitability of the modeled phenomenon in underdeveloped theories, that lack proper concepts. Here, modeling acts as a surrogate for experimentation, eventually leading to conceptual improvements. Thus, it is also worth emphasizing in this text the use of *scientific modeling for exploration*, in the context of deep learning.

It should be evident by now that the "neurons" we have been talking about in this introductory text pay almost no resemblance to the biological ones. However, given that many deep learning models draw inspiration from animal neural systems (humans included) for the sake of their architectures, I wonder whether their stunning successes in tasks like object recognition, planning and language processing, among others, may somewhat reflect, despite the oversimplified nature of the artificial neurons themselves, some yet unknown general structural and functional principles. This is, I submit, yet another productive pathway open to investigation. Considering the biological side of this topic, a good starting point for the philosopher interested in this line of research is, in my opinion, the book by Sterling & Laughlin (2015).

I note that further exploration of this interesting theme would lead us too far away from my original goal in this paper. Once said that, let us begin to address, in the next section, the negative aspects of the technologies under scrutiny. In the next sections, we are going to sequentially examine three problems that arose from the wide adoption of the technologies mentioned above: misuses of big data, deepfakes (and similar abuses of technology), and the black box dilemma in deep neural networks.

3. Big Data

The fast pace of the recent technological development observed in the computer industry (faster processors, larger data storage equipment, and fast communications) allowed for the creation of datasets that are so big, that they represent a real challenge to traditional database management systems. The term "big data" arose in this context, where a single dataset may sometimes contain many petabytes (one petabyte is equivalent to 1000 terabytes) of data (Elgendy & Elragal 2014).

There is, however, some disagreement about the exact meaning of "big data." Indeed, some authors point out that the term refers not only to size (as mentioned above), but also to several computational methods used in dataset analyses. Curiously enough, even though big data appears in a growing body of philosophical literature that addresses the role of modeling and simulation (and, more generally, the role of software in scientific endeavors), the term is rarely seen in texts of philosophy of science (Symons & Alvarado 2016).

According to Leonelli (2020), the big data phenomenon is clearly the outcome of the "datafication" of society, where all sorts of human actions are registered and stored, yielding a huge and ever-growing digital footprint. New ways to produce, store and subsequently analyze data led to the establishment of the field of *data science*, where knowledge is extracted from big data by means of mathematical, statistical, and algorithmic/computational techniques and tools. The *Open Data* movement played an important part here, promoting the sharing (and interlinking) of diverse types of research data by means of digital infrastructures. A huge amount of readily available machine-readable formatted data has stimulated the development of efficient processes that collect, organize, present and model it, leading, ultimately, to the production of ever more powerful platforms of artificial intelligence. Interestingly enough, researchers from essentially all areas of inquiry are starting to use these new data manipulation capabilities to improve their own scientific efforts (Leonelli 2020).

Symons & Alvarado (2016) pointed out that what characterizes and distinguishes the mainstream take on big data from other approaches is its combined use of statistical methods and computational analytical tools. According to Leonelli (2020), pertinent examples of such tools and methods are machine learning tools and, more specifically, deep neural networks (among other "intelligent" data handling practices). In this field, Leonelli goes on, expertise in programming and even in computer hardware architectures (as in computer engineering) is often necessary, in addition to more traditional mathematical and statistical abilities. This amounts to a somewhat different epistemological approach to scientific research, namely a data-centric one, that places more emphasis on the processes by means of which the research is conducted than on the research outcomes themselves.

I would like to emphasize that a large part of this kind of research focuses exclu-

sively on the development of tools for commercial purposes, hence it usually takes place in fancy research labs of big companies like Google, Facebook, Amazon, Microsoft, and Uber, to name but a few. This line of research is very expensive, due to the need for high-end computer hardware for deep neural networks training and such, and is, therefore, usually restricted to a few selected groups of researchers. To be fair, these companies also tend to provide, as a by-product, plenty of inexpensive or even free computational resources for everyone interested. Popular resources, such as Kaggle and Google Colaboratory, provide many hours of free GPU cloud computing for developing and training deep neural networks and other machine learning models.

There are certainly potential risks in *applying* deep neural networks to all sorts of problems, such as autonomously driving vehicles, and perhaps enabling the next wave of autonomous robotic weapons. But what about *training*-related risks? It is widely known that most deep neural networks (and other machine learning models) require huge amounts of training data, in order to perform optimally. Hence, one wonders if there are risks *specifically* associated with collecting large amounts of data with the main purpose of training deep learning systems and related data models. If so, it is important to know exactly (Leonelli 2020) the nature of the risks, who can be harmed in the process and, hopefully, what can be done in order to minimize the possibility of harsh consequences taking place.

Let us focus, then, on some of the risks associated with collecting big data, as well as related ethical issues, as highlighted by Leonelli (2020). A first point of concern is that some kinds of data, e.g., personal data, are collected, curated, and commercialized by big corporations for all sorts of purposes, including scientific research (e.g., for training machine learning models), thus creating a gap between those who can and those who cannot (or maybe do not want to) pay for accessing and using the data. Besides, it is certainly the case that these corporations release data selected according to their own interests, creating thus unavoidable distortions and biases. This is the case, e.g., of data that they cannot interpret alone, needing assistance, in some degree, from the public sector (and this may result in some forms of exploitation), as well as of data that they consider less profitable anyway. In the end, says Leonelli, it is pretty clear that the economic value of big data tends to overwhelm its scientific value.

As far as biases are concerned, there are unfortunately plenty of examples around. As pointed out by Buolamwini & Gebru (2018), a number of machine learning models for human face recognition are trained with labeled data, and if the data is biased, this results in "algorithmic discrimination." These authors studied three commercial systems for gender classification and showed that the accuracy in identifying lighter-skinned males (maximum error rate of only 0.8%) is much larger than the accuracy in identifying darker-skinned females (maximum error rate of 34.7%). Obermeyer

et al. (2019), by their turn, showed that a given commercial prediction algorithm, widely used in US hospitals, has been highly discriminative, being *less likely to refer black people* than white people, equally sick, to health programs that improve care for patients demanding complex medical needs.

Another point that demands attention (Leonelli 2020) is the growing number of databases that cease to be actively maintained due to lack of funding or other reason. This fact may be hidden from its potential users, that may thus be unknowingly dealing with obsolete data. Leonelli (2018) offers the example of a given fungi database, containing data collected decades ago, that is most definitely unreliable to explain the behavior of the same species of fungi nowadays (or in the future). The uninformed use of such database may certainly seriously compromise any scientific analysis.

Still according to Leonelli (2020, p. 40), "it is essential that ethical and social issues are seen as a core part of the technical and scientific requirements associated with data management and analysis." It is important to note that regulations of the commercialization of research, or of how personal data are handled, are very welcomed, albeit insufficient, measures: "To guarantee that big data are used in the most scientifically and socially forward-thinking way it is necessary to transcend the concept of ethics as something external and alien to research" (Leonelli 2020, p. 40). That is to say, all those who manage the data (and the methods used to visualize and analyze it) should ideally go beyond the purely "technical," and effectively internalize the ethical concerns, ensuring thus an ethically sound management of data.

As Leonelli (2020) points out, it is clear that many contemporary scientists view their own *efforts* to deal with huge amounts of data as fundamental contributions to the endeavors of science. This data-centric approach to scientific discovery, however, signals the enormous challenges ahead, concerning gathering, storing, organizing, classifying, analyzing, interpreting, and finally understanding data. It is worrisome the rate at which data is being produced by apps running in smart devices, the same being true of specially designed high-throughput instruments measuring all sorts of variables (in medicine, industry and so on). Even though the data is in principle passive of fast delivery to everyone concerned, there are many technical, legal, and ethical consequences of concern. Besides, even when access is granted, the usefulness of the data for the sake of research is not guaranteed. Finally, it could be argued that the opaque nature of many machine learning tools may compromise the scientific meaning and credibility of the fruits of a given scientific initiative, but this is a controversial topic. Leonelli also mentions the fact that some philosophers already wonder if mankind risks eventually losing control of the whole scientific (and technological) enterprise to some sort of foreign, artificial, complex, and inscrutable intelligence.

Leonelli (2020) also reminds us that much of contemporary philosophy of science is predicated upon an understanding of rationality grounded on individual agency,

as well as on the display of a whole spectrum of cognitive capabilities. It could be said, perhaps, that big data science drifts away from these long-cherished ideals at a fast pace, but this is open to discussion. Besides, the decentralized, interlinked, and distributed nature of data production and dissemination channels, as well as of big data infrastructure and analytics-related decision making, require the skills of many people all over the world, so that no individual holds the power to directly influence outcomes. Hence, Leonelli goes on, it follows that the analysis of big data may be the ultimate distributed cognitive system.

4. Deepfakes

We have already seen hints of unintentional misuses of deep learning models, mostly related to big data. Of great concern as well is, as mentioned, the intentional employment of machine learning methods and technologies in general (not only deep learning models) in certain areas. A drastic example is the use of artificial intelligence in "smart weapons," be it in military autonomous weaponized (aerial and terrestrial) robots, robotic machine guns, fully autonomous "submarine-hunter" submarines, or other applications already in use. There is yet another intentional misuse of deep neural networks that, I believe, deserves a closer look, namely the appropriation of this kind of technology for deliberately producing the so-called *deepfakes*.

One way of defining a "deepfake" is to consider it as being a digital medium (image, video, recorded speech, etc.) containing swapped person identities. According to Somers (2020), the term was coined by a user of the Reddit platform who opened, in late 2017, a virtual space for exclusively sharing videos of pornographic nature that were altered using face-swapping open-source code. In the following year, a deepfake video portraying the former American president Barack Obama was released, raising widespread concerns about identity theft, impersonation, and the dissemination of false information through social media (Mirsky & Lee 2021). Nowadays the term has a wider meaning, incorporating both, the products of techniques that were already in use before the term was conceived, and results of new and more efficient techniques, e.g., the use of StyleGAN (Somers 2020). StyleGAN is an interesting example of yet another kind of deep learning model, namely the so-called "generative adversarial networks."

A *generative adversarial network* (GAN) consists of two deep neural networks combined, namely a *generator* network and a *discriminator* one (Russell & Norwig 2020). The generator competes against the discriminator in the context of a game-theoretic scenario, where the generator produces fake "samples," while the discriminator tries to distinguish between the generated samples and those real samples drawn from the training dataset. The scenario is said to be game-theoretic because

the competition between both neural networks is in the form of a "zero-sum game," a classic situation studied in the discipline of game theory, in which the gain for one agent means a loss for the other, so that the game's net improvement in benefit is null.

During training, the generator attempts to make the discriminator believe its fake samples are real, while the discriminator attempts to correctly classify each sample into the real or fake categories. When the training process ends, the fake samples are virtually indistinguishable from the real ones (Goodfellow; Bengio; Courville 2016). At this point, the discriminator has completed its task, and goes straight into the trash bin, as a manner of saying. The style-based generative adversarial network architecture (StyleGAN) is a specific kind of generative adversarial network, developed by NVIDIA researchers (and open-sourced at the beginning of 2019), that generates synthetic human faces (Karras; Laine; Aila 2019; Karras et al. 2020). This kind of model can be misused to produce very convincing deepfakes.

There are formidable ethical challenges ahead, regarding the indiscriminate employment of deepfakes (de Ruiter 2021). An area of great concern, due to the widespread availability of this kind of technology, is the so-called *nonconsensual deepfake porn*, where non-intimate images of someone, usually a woman, are taken from social media accounts (or other sources available at the internet), manipulated by deep neural networks to fit into a pornographic photo or video scene, and subsequently released to the public. This is frequently associated with a phenomenon known as "revenge porn." Fortunately, this kind of technology-enabled abuse is starting to grab the attention of lawmakers (up to now, mainly in the UK and the USA), due to strong pressure from activists (Hao 2021). Other misuses of deepfakes include, but are not restricted to: blackmailing, sabotaging, intimidating, spreading different kinds of misinformation, impersonating personalities (political figures, musicians, actors and actresses, and so on), promoting violence, ideologically influencing, and defaming chosen individuals in all sorts of manners beyond nonconsensual deepfake porn (de Ruiter 2021; Mirsky & Lee 2021).

Apart from the investigation of the ethic aspects surrounding the proliferation of deepfakes, there is certainly more to be explored here, making this a good choice for someone looking for a fresh line of research. An example is the issue of the "epistemic threat" of deepfakes, where it is argued by Fallis (2021) that deepfakes restrict the information carried by videos, thus damaging the viewers. Another example is the extent to which deepfakes may, in the near future, seriously compromise the reliability of our testimonial practices, by degrading the epistemic value of recordings (Rini 2020). To cite but one more example, Hancock & Bailenson (2021) explicitly emphasize the need for more researchers to begin studying the myriad of social issues that surround the employment of deepfake technology. In the next section, we are going to address some important aspects of the "black box" problem in artificial

intelligence, with special emphasis to deep learning.

5. Black Boxes

Even though the quality and resolution of images created by generative adversarial networks (also known as GANs, see above) have improved at a fast pace, the fact is that these models are still largely huge black boxes, hindering the understanding of many aspects of the image synthesis process (Karras; Laine; Aila 2019). Generative adversarial networks are very successful and, in the opinion of Russell & Norwig (2020, p.750), "the true reasons for the success of deep learning have yet to be fully elucidated." When discussing opacity in machine learning systems such as deep neural networks, it is useful to distinguish between two concepts, namely interpretability and explainability.

As explained by Russel & Norwig (2020), a machine learning model is said to be *interpretable* if, when inspecting the model itself, one understands *why* a certain output was provided by the model for a given input (and *how* an input change would result in a different output). Examples of highly interpretable models are the so-called decision tree and linear regression models. A machine learning model is said to be *explainable*, on the other hand, *if the model itself or a separate process helps one understand why* a certain output was provided for a given input. As mentioned, explainability can be provided by another process, so that even an uninterpretable (black box) model can still be explained by means of a summary provided by an explanation module. Explanation modules effectively treat a machine learning model as a black box, providing different inputs to create a dataset from which to build the new interpretable model, that is usually a linear regression or a decision tree model (that by their turn are, as mentioned above, highly interpretable). The new interpretable model is thus an *approximation* of the original uninterpretable one and can be used to create explanations about the relative importance of each feature. Unfortunately, this approach, that works well with structured data, fails for data like images, where each and every pixel is a feature, but nonetheless lacks "importance" (Russell & Norwig 2020).

One may prefer a given model class over another simply because it is explainable, thus eventually choosing "trust" over accuracy. But one needs to pause for a moment and wonder if this sense of security is justified (Russell & Norwig 2020). Here, the authors raise a valid point: if the machine learning model is to be constructed with the sole purpose of helping clarify a given domain of knowledge, then interpretability and explainability are most certainly fundamental criteria to be pursued. If, however, the goal of the prospective machine learning model is simply to yield top performance in a given technological context, then extensive testing of the

model in working conditions is far more reassuring than detailed explanations. One wonders if it is better to travel in an undocumented self-driving car that has already impeccably performed hundreds of trips around the town without a single incident, or in a new car that just came out of the factory, but that is accompanied by a manual containing detailed schematics and explanations endorsing the safety of all its artificial intelligence subsystems.

To close this section, I would like to emphasize that the investigation of the interpretability and explainability of opaque deep learning models has been gaining momentum in recent years (Lipton 2018; Montavon; Samek; Müller 2018; Park et al. 2018; Zednik 2019; Erasmus; Brunet; Fisher 2021). This is a rich source of material for philosophers that are new to this important subject matter, but that are nonetheless willing to jump into action as soon as possible.

6. Final Remarks

As stated at the introduction, the main goal of this text was to provide an overview of these topics to new researchers that are interested in this important area of model application. Accordingly, I initially mentioned some important chronological facts, presented essential definitions (of artificial intelligence, machine learning, deep learning, and so on), and examined noteworthy aspects related to the field of deep neural network models and its applications. After that, I sequentially addressed three problems of note, all of them directly related to deep learning: big data and its shortcomings, deepfake pandemics and the opacity of complex machine learning models.

We saw that the problems of big data are multifaceted, starting with its own definition, that is still debated now and then. We also became acquainted with some of its deep social implications, that include important issues such as strong control by large corporations, information filtering, and the monetization of individual data. And there are also the issues of obsolete databases, biased data, and algorithmic discrimination, as well as public health access inequalities. But we also noticed that there are some ways of improving the situation.

Apart from big data, there are issues pertinent to other explicit misuses of deep learning, such as deepfaking, i.e., the conscious swap of person identities with the probable goal of harming, engaging in defamation, or taking revenge. And as we saw, one appalling manifestation of such regrettable practices is nonconsensual deepfake porn, made possible thanks to StyleGAN and other deep neural network models. But we also learned that things are seemingly changing for the better, thanks largely to the active role played by activists, and most possibly other social actors as well.

The last problem addressed in this text, which admittedly received less space than the others, was the issue of black boxes, or the fact that many machine learning

models, deep neural network models included, are opaque to internal scrutiny. We examined the notions of interpretability and explainability of models, but we also learned that "blind" industrial-strength performance of a given deep learning model, in an economic and technological context, is sometimes preferable than a full understanding of its entrails.

In this paper I drew attention to some uses of modeling found in the literature, associated to artificial intelligence (with emphasis in deep learning), namely: (i) data modeling, (ii) scientific modeling of cognition and (iii) scientific exploration via models. The latter approach, as we saw, may result, according to its proponents, in new hypotheses, proof-of-principle demonstrations, and "suitability checking" in underdeveloped theories. All these uses, I suggest, deserve a closer look, and may well be a starting point for yet another philosophical project.

In the not-so-distant past, philosophers engaged in heated debates about artificial intelligence and artificial neural networks. Strangely enough, now that deep learning models clearly reached another level of performance, being capable of doing things previously thought impossible (and irreversibly impacting science and society in fundamental ways), there are relatively few philosophers dealing with this subject matter. This void is even more apparent when one considers that the huge successes of the new data models remain largely unexplained, as we saw.

References

Arbesman, S. 2013. Five Myths About Big Data. *The Washington Post*. http://www.washingtonpost.com/opinions/five-myths-about-big-data/2013/08/15/64a0dd0a-e044-11e2-963a-72d740e88c12_story. Access: 29.03.2021.

Berggren, K. et al. 2020. Roadmap on Emerging Hardware and Technology for Machine Learning. *Nanotechnology* **32**(1): 012002. https://doi.org/10.1088/1361-6528/aba70f. Access: 03.05.2021.

Bernard, E. 2021. *Introduction to Machine Learning*. Champaign: Wolfram Media, Inc.

Boyd, D. & Crawford, K. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society* **15**(5): 662–679.

Buolamwini, J. & Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research* **81**: 1–15.

Cichy, R. & Kaiser, D. 2019. Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences* **23**(4): 305–317.

de Ruiter, A. 2021. The Distinct Wrong of Deepfakes. *Philosophy & Technology* **34**: 1311–1332.

Elgendy, N. & Elragal, A. 2014. Big Data Analytics: A Literature Review Paper. In: P. Perner (ed.), *Advances in Data Mining: Applications and Theoretical Aspects*, p. 214–227. Cham: Springer.

Erasmus, A.; Brunet, T. D. P.; Fisher, E. 2021 What is Interpretability? *Philosophy & Technology* **34**: 833–862.

Fallis, D. 2021. The Epistemic Threat of Deepfakes. *Philosophy & Technology* **34**: 623–643.

Frankish, K. & Ramsey, W. M. 2014. Introduction. In: K. Frankish & W. M. Ramsey (ed.), *The Cambridge Handbook of Artificial Intelligence*, p.1–11. Cambridge: Cambridge University Press

Goodfellow, I.; Bengio, Y.; Courville, A. 2016. *Deep Learning*. Cambridge, MA: MIT Press.

Hancock, J. T. & Bailenson, J. N. 2021. The Social Impact of Deepfakes. *Cyberpsychology, Behavior, and Social Networking* **24**(3): 149–152.

Hao, K. 2021. Deepfake Porn is Ruining Women's Lives. Now the Law may Finally Ban it. *MIT Technology Review*.
https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/. Access: 22.07.2021.

Haykin, S. 2009. *Neural Networks and Learning Machines*. 3rd Edition. New Jersey: Pearson.

Kandel, E.R.; Barres, B. A.; Hudspeth, A. J. 2013. Nerve Cells, Neural Circuitry, and Behavior. In: E. R. Kandel; J. H. Schwartz; T. M. Jessell; S. A. Siegelbaum; A. J. Hudspeth (ed.), *Principles of Neural Science*, p. 21–38. 5th Edition. New York: McGraw Hill.

Karras, T.; Laine, S.; Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 4396–4405.

Karras, T.; Laine, S.; Aittala, M; Hellsten, J.; Lehtinen, J; Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 8107–8116.

Leonelli, S. 2018. The Time of Data: Timescales of Data Use in the Life Sciences. *Philosophy of Science* **85**(5): 741–754.

Leonelli, S. 2020. Scientific Research and Big Data. In: E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Summer 2021 Edition. https://plato.stanford.edu/archives/sum2021/entries/science-big-data/. Access: 31.07.2021.

Lipton, Z. C. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is both Important and Slippery. *acm queue* **16**(3): 1–27.

McCulloch, W. S. & Pitts, W. H. 1943. A Logical Calculus of the Ideas Immanent in Nervous Activity. *Bulletin of Mathematical Biophysics* 5: 115–133.

Mirsky, Y. & Lee, W. 2021. The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys* **54**(1): 1–41.

Montavon, G.; Samek, W.; Müller, K. R. 2018. Methods for Interpreting and Understanding Deep Neural Networks. *Digital Signal Processing: A Review Journal* 73: 1–15.

Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. 2019. Dissecting Racial Bias in an Algorithm used to Manage the Health of Populations. *Science* 366: 447–453.

Park, D. H.; Hendricks, L. A.; Akata, Z.; Rohrbach, A.; Schiele, B.; Darrell, T.; Rohrbach, M. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 8779–8788.

Rini, R. 2020. Deepfakes and the Epistemic Backstop. *Philosopher's Imprint* **20**(24): 1–16.

Russell, S. & Norvig, P. 2020. *Artificial Intelligence: A Modern Approach*. 4th Edition. Boston: Pearson.

Seung S. & Yuste, R. 2013. Neural Networks. In: E. R. Kandel; J. H. Schwartz; T. M. Jessell; S. A. Siegelbaum; A. J. Hudspeth (ed.), *Principles of Neural Science*, p. 1581–1600. 5th Edition. New York: McGraw Hill.

Somers, M. 2020. Deepfakes, Explained. *MIT Sloan School of Management*. https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained/. Access: 28.07.2021.

Sterling, P. & Laughlin, S. 2015. *Principles of Neural Design*. Massachusetts: MIT Press.

Symons, J. & Alvarado, R. 2016. Can we Trust Big Data? Applying Philosophy of Science to Software. *Big Data & Society* **3**(2): 1–17.

Thagard, P. 1990. Philosophy and Machine Learning. *Canadian Journal of Philosophy* **20**(2): 261–276.

van Veen, F. & Leijnen, S. 2016. A Mostly Complete Chart of Neural Networks. https://www.asimovinstitute.org/neural-network-zoo/. Access: 30.09.2021.

Zednik, C. 2019. Solving the Black Box problem: a normative framework for explainable artificial intelligence. *Philosophy & Technology* **34**: 265–288.