AI & SOCIETY (2020) 35:309–317
https://doi.org/10.1007/s00146-019-00888-w

ORIGINAL ARTICLE

Black-box artificial intelligence: an epistemological and critical analysis

Manuel Carabantes[1]

Received: 3 February 2019 / Accepted: 27 March 2019 / Published online: 12 April 2019

Abstract
The artificial intelligence models with machine learning that exhibit the best predictive accuracy, and therefore, the most powerful ones, are, paradoxically, those with the most opaque black-box architectures. At the same time, the unstoppable computerization of advanced industrial societies demands the use of these machines in a growing number of domains. The conjunction of both phenomena gives rise to a control problem on AI that in this paper we analyze by dividing the issue into two. First, we carry out an epistemological examination of the AI's opacity in light of the latest techniques to remedy it. And second, we evaluate the rationality of delegating tasks in opaque agents.

Keywords Artificial intelligence · Philosophy of technology · Machine learning · XAI · Deep neural networks · GDPR · Instrumental reason

1 Introduction

At 14:32 on May 6, 2010, a flash crash occurred on the stock exchange in the United States—a quick fall in security prices. In a matter of minutes, the three most important stock market indices in the country (S&P 500, DJIA and Nasdaq Composite) sank and rose again, registering variations of approximately 1000 points (CFTC & SEC 2010) that implied the momentary loss of trillions of dollars. Beyond the human errors, which began to be debugged in 2015, the phenomenon had a technical explanation: the computers in charge of carrying out stock transactions did not work as their programmers had expected. The automatic trading systems (ATSs) went into a spiral of buying and selling that distorted stock prices, until a few minutes later, at 14:45, an automatic security mechanism stopped the process. The earthquake had passed. It was time to help the victims and understand what had happened to prevent it from happening again.

To reverse the damage caused as much as possible, representatives of the exchanges and regulators met after the session to, in compliance with the legislation, cancel the transactions carried out at prices too distant to those before the crisis. Regarding the understanding of the phenomenon to prevent it from happening again, the engineers in charge of the ATSs found an explanation in the memory registers: the computers had applied the action rules they were programmed with, but they had done it in an unforeseen context in which these rules, which usually produce desired results, were catastrophic (Bostrom 2014). This is a problem that has been known in Artificial Intelligence (AI) for decades as the qualification problem (McCarthy and Hayes 1969). Therefore, from the technical point of view, nothing new happened under the sun. But, what if the engineers had not been able to discover the cause of the abnormal behavior of computers? What if the ATSs had architectures that made it impossible to understand their decisions? Could the stock exchange market have opened the next day?

In the field of artificial intelligence, these machines exist: they are opaque or black-box models that return outputs, or that make decisions by applying a decision theory on the outputs, by running internal processes that are incomprehensible by human beings. Moreover, they are not marginal models but widely used. As a recent paper by the U.S. agency DARPA points out, within computers with machine learning (ML), that is, capable of learning from experience, those with the greatest predictive power are those with the most opaque architectures: "There is an inherent

✉ Manuel Carabantes
manuel.carabantes@gmail.com

[1] Faculty of Philosophy, Universidad Complutense de Madrid (Complutense University of Madrid), c/ Profesor Aranguren 5, 28040 Madrid, Spain

tension between machine learning performance (predictive accuracy) and explainability; often the highest performing methods (e.g., deep learning) are the least explainable, and the most explainable (e.g., decision trees) are less accurate" (DARPA 2016, 7). In an increasingly computerized world in which not only the stock market economy, but also defense, energy, communications, transportation, and even food depend on artificial intelligence, is it reasonable to trust AI models that are not very comprehensible just for the sake of their greater effectiveness?

This paper will answer this question through the following itinerary. First, we will perform an epistemological analysis of the opacity in the three forms in which it appears in the ML field according to Jenna Burrell: opacity as intentional concealment, opacity as technological illiteracy and opacity as cognitive mismatch (Burrell 2016). Opacity as cognitive mismatch is the most worrisome, since it also prevents the engineers who develop certain ML models to understand how their own creations work. That is why we will dedicate a section to expose the techniques that have emerged in recent years to remedy it. And, finally, in the section dedicated to the conclusions, we will use Horkheimer's critique of instrumental reason to answer the question about the rationality of delegating tasks in artificial agents with black-box architectures.

Perhaps a preliminary clarification on the use of the term epistemology is necessary. As the reader might know, epistemology is a usual label to refer to the theory of knowledge. We rightly say that our analysis is epistemological with regard to opacity as intentional concealment and technological illiteracy. But on opacity as cognitive mismatch the issue is more properly about cognition. So we are using epistemology in a broad sense that includes the knowledge problem at many levels: from the human interaction, in which interlucency is a key concept that we will bring, to the individual facing a phenomenon that does not fit his cognitive forms, which is what happens when we try to understand the functioning of an artificial neural network. Maybe a good alternative would be talking about gnoseology, as it is a term that embraces the first two opacity forms, and at the same time is etymologically linked to cognition. But we should keep in mind that both terms are commonly used as synonyms, and in Anglo-Saxon philosophy epistemology has been often the only term used (Ferrater 2009). So, as this paper is in English, and after this preliminar statement, we consider that it is fair to entitle our work as an epistemological analysis.

2 Machine learning

Machine learning (ML) is the branch of artificial intelligence that aims to develop techniques for making computers learn. "An agent learns if it improves its performance on future tasks after making observations about the world" (Russell and Norvig 2016, 693; Benítez 2013). In recent years, ML has become an essential tool for a broader spectrum of applications such as image classification, speech recognition, and natural language processing (Montavon et al. 2017). In addition, ML systems have proven to be better than human experts in strategic games such as Go, poker, and chess (Samek et al. 2017). In daily life there are many classification and evaluation mechanisms with social impact that ultimately rely on ML: spam filters, credit card fraud detection, search engines, news trends, market segmentation and advertising, insurance or loan qualification, and credit scoring (Burrell 2016). Even without being aware of it, millions of decisions per second that affect our lives are made by computer systems equipped with ML, and the tendency is to increase that delegation of tasks in computers.

For our interests, we can divide ML techniques into two sets: those that allow us to understand the operation of the machine (non-opaque), and those that do not (opaque). In the first group would be the linear methods, and in the second, the non-linear ones, among which the artificial neural networks (ANN), inspired by the functioning of the nervous system of living beings (Rumelhart et al. 1989), stand out for their effectiveness. As we have indicated before, the reality today is that the most effective ML techniques are also the most opaque ones (DARPA 2016). This is precisely our object of study: opaque ML techniques. And within them we will focus on deep neural networks (DNN), a subtype of ANN that is defined by having more than one hidden layer, giving them a greater capacity of internal representation (a problematic concept to be addressed elsewhere). The content of this paper will be applicable in certain cases to support vector machines (SVM) and other ML techniques, but due to the need to delimit the scope of our exam, and because DNNs are the most effective ML models and at the same time the most opaque ones, they will be our main focus of attention.

The fact that ANNs are problematic because of their opacity is something that has been known for decades. We can give examples ranging from the anecdote remembered by Hubert Dreyfus about one of the first ANNs that was supposed to recognize tanks when in reality it had learned to differentiate between sunny and cloudy days (Dreyfus 1992, xxxvi), until the recent discovery of what two models for recognizing horses had surprisingly learned (Fraunhofer 2017). One was a Fisher vector (FV) classifier, and the other one was a DNN. Both exhibited similar success rates in the recognition of images that contained horses. However, researchers found out that the FV knew nothing about horses: what it had learned to recognize were the copyright symbols that referred to websites about horses (Lapuschkin et al. 2016). The discovery of the error was possible thanks to the LRP, one of the state-of-the-art techniques proposed

to explain the decisions of DNNs and that, however, are still far from solving the opacity problem.

3 Forms of opacity

So far we have talked about opacity only in one of its forms—opacity as lack of comprehensibility of ML models. It is an issue that we are going to examine further, including the exposition of techniques to remedy it, such as the LRP. But before going into this form of opacity it is necessary to draw a topology of the concept relative to the ML field. Burrell's (2016) is pretty good. According to it, there are three forms of opacity: (1) opacity as intentional corporate or institutional concealment of their algorithms in general, including those of ML; (2) opacity as technological illiteracy that prevents society from understanding a field as specialized as that of computer programming; and (3) opacity in the sense presented so far, that is, opacity as cognitive mismatch between the complex mathematical operations performed by ML algorithms and the type of reasoning used by human beings. However, Burrell's text has shortcomings that require to review and extend its content in light of what happened in recent years in computer science and beyond.

3.1 Opacity as intentional concealment

Large companies and governments hide the algorithms they use to process data, including the highly effective ML predictors and classifiers. They argue that such discretion is necessary for two main reasons: security and competitiveness. In terms of security, revealing the code of their computer programs would make them more vulnerable to malicious attacks. And with respect to competitiveness, companies have the right to industrial secrecy as a basic game rule to be competitive in the free market economy. However, for some technologies discretion is much more than a factor of competitiveness: it is inherent to their proper functioning. For example, if Google, which has ML among its subsystems, revealed its website classification algorithm, then search engine optimization (SEO) experts would use that information to carry out exploits, that is, an advantage to get unwanted behavior on the part of the search engine. It is Odysseus' guile: to comply with the norm without complying with the spirit beneath it. Just as Odysseus complies with the obligation to listen to the sirens' song, but by ordering his sailors to tie him to the mast he mocks the spirit of the mandate—which is to get him dead by jumping into the sea—the details of how Google classifies websites would serve SEO experts to comply with the demands of Google to obtain a good website classification but without complying with the spirit of those demands, which is, for the web be a relevant output for certain keywords.

Burrell takes into account the criticism of Frank Pasquale on this point: companies and governments may be using these arguments as subterfuges to conceal that their algorithms violate legality or morality—implementing discriminatory practices for example (Pasquale 2015). The solution could either be direct or indirect. The direct one would be the enactment of laws that allow independent commissions to examine the source code of the algorithms (Diakopoulos 2013), whereas the indirect solution would be to explore the algorithms from the outside, with or without permission (Sandvig et al. 2014).

As Bryce Goodman and Seth Flaxman point out, the general data protection regulation (GDPR) comes to terminate, to some extent and directly, this form of opacity as intentional concealment in the European Union (Goodman and Flaxman 2016). But to what extent? Little. Approved by the EU Parliament on April 14, 2016 and enforced on May 25, 2018, the GDPR is familiar to residents in the EU, it is the cause of whenever we surf the Internet we are continually assaulted by banners asking us to accept the cookie policy of each website, because it is one of its imperatives to inform users that their information will be processed computationally. On the opacity of the algorithms used by companies and governments, the GDPR states in Article 13(2) "The controller shall, at the time when personal data are obtained, provide the data subject with the following further information necessary to ensure fair and transparent processing: […] (f) The existence of automated decision-making, including profiling, referred to in Article 22(1) and (4), at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject". When our data are part of an automated information processing, we must be informed of the logic involved. What does this mean? Little to nothing. As Goodman and Flaxman point out, "It is not clear whether companies will be required to disclose their learning algorithms or training datasets and, if so, whether that information will be made public" (Goodman and Flaxman 2016). It was not clear in 2016, and it still is not in 2019.

However, even assuming that companies and governments were obliged by the EU to the greatest possible transparency, which would consist in the disclosure of the algorithms' source code, the GDPR would not be enough to eliminate opacity as concealment due to the fact that it considers epistemic generosity (i.e., informing the data subject) as if it were non-epistemic generosity. Other-regarding epistemic virtues, such as giving information, differ from non-epistemic virtues, according to Boudewijn De Bruin, in that they require the recipient to cooperate (De Bruin 2015). In addition, there must be active tracking by the sender to ensure that the recipient understood the message. When these two requirements are fulfilled, as De Bruin says, then

interlucency occurs (De Bruin and Floridi 2016). It is true that the GDPR states the right to be attended by a human being [Art. 22 (3)], and to lodge a complaint with a supervisory authority [Art. 13 (2.d)]; through this way a process of interlucent dialogue could begin in the best of cases. But the reality is that, on the one hand, nobody reads the cookie policy that web pages are obliged to display, so the requirement of the recipient's cooperation is missing; and on the other hand, those who process our data usually offer only unilateral communications, without making sure that we understand the legal or significant ramifications on our lives [Art. 22 (1)] that their activity may have for us, so the active tracking of the recipient's understanding by the sender is also missing. The GDPR does not guarantee interlucency; and without interlucency, epistemic generosity is not authentic generosity, but mere imposture.

As a working hypothesis, let us suppose that future jurisprudence and legislation set a maximal interpretation of the GDPR regarding the informative obligations of the data processors, and that both sender and recipient of information engage in an interlucent dialogue. Even in this idealized scenario, we should still overcome the other two forms of opacity, starting with technological illiteracy. Because, of what use would be the disclosure of ML algorithms used by large companies and governments to a subject who, although interested in them as interlucent, does not know anything about computer programming?

3.2 Opacity as technological illiteracy

Burrell rightly points out that reading and writing computer code are highly specialized skills that the majority of the population lacks (Burrell 2016). However, she does not succeed in the solutions proposed to tackle it. They are, to put it academically, very innocent, naive, typical of what Gustavo Bueno referring to Lewis Carroll's character called the Alice thinking style (Bueno 2006). The first, taken from Nicholas Diakopoulos, is that journalists transmit that specialized information in a more affordable format, just as they do when reporting on politics, economy, or laws (Diakopoulos 2013). The second is to intensify efforts in the education system to instruct the next generations in STEM fields (Science, Technology, Engineering and Mathematics). Let us examine both.

The hypothetical success of mass media in disseminating knowledge about computers can be predicted by evaluating the success they had in spreading knowledge on other areas they already deal with. In *Democracy and political ignorance*, Ilya Somin collects various statistical data from the last half century that reflect how American media, paradigm for acts of bravery such as the Watergate or the Pentagon Papers, has failed to enlighten the Americans even in the most elementary issues of public life (Somin 2013). For example, in 1964, during the Cold War, only 38% of the population was aware that the Soviet Union was not a part of NATO, a military alliance led by the U.S. to face the communist threat. A survey taken after the 2002 congressional elections revealed that only 32% of respondents knew that the Republicans held control of the House of Representatives before the election. In 2006, a study showed that only 42% knew that the three branches of the federal government are executive, legislative, and judicial. Like these, we could cite many other figures that support our distrust on the idea that journalism could be the remedy to opacity as technological illiteracy. If with all the attention of the media on politics, only 32% knows which party controls the Congress before an election, how many headlines and hours of news on TV should be devoted to the generalized delta rule so that a similar percentage could understand this key concept in the functioning of ANNs?

Regarding education, Burrell does not take into account the complex nature of modern technique. Ortega y Gasset describes this phenomenon as follows: "The feudal lord, for example, saw his horses shod, his lands plowed, and his wheat grinded. Today he not only does not usually see the corresponding techniques in action, but most of them are invisible. I mean that looking at them does not reveal their reality, does not make them intelligible. Seeing a factory may leave an aesthetic, emotive impression, but it does not teach congruently what is the technique of that factory, like seeing a car does not disclose the complicated plan of its machinery" (Ortega y Gasset 2009, 29). The capitalist, who is the equivalent to the feudal lord, does not know the functioning of the ATSs that make him rich in the stock exchange. Clearly, the capitalist knows the rudiments of operation at the level at which they were described at the meeting with the financial analysts to decide the market strategy; but nothing more. An ATS with ML can use the generalized delta rule, and the capitalist who owns it does not know what it is. Modern technique is not like the medieval or the ancient one; it is usually invisible, and it is rare that human beings experience the Aristotelian wonder and become interested in what is hidden from the gaze.

What is the cause of the ignorance of the technique, both by the citizen and by the owner who is enriched by it? Jason Brennan would say that it is not stupidity, but what economists call rational ignorance: "When the expected costs of acquiring information of a particular sort exceed the expected benefits of possessing that sort of information, people will usually not bother to acquire the information" (Brennan 2016, 30). Against this wall also clashes interlucency, of which we have discussed in the previous section, it is not reasonable to believe that the user will spend time reading the cookie policy when he does not foresee that this action will generate a return that rewards the effort. Technological illiteracy covers and will cover all domains

of computers that do not produce the kind of rewards that ordinary people aspire to, which are by the way, very far off from the spiritual pleasures that John Stuart Mill said distinguish us from the beasts.

However, let us again grant as a working hypothesis that opacity as technological illiteracy is remediable, either through a new honest journalism ethic committed to the truth over the interests of the advertisers, or thanks to a new revolutionary school capable to penetrate to the depths of the limbic system to educate human beings in loving wisdom. Would that be enough to remedy the opacity of ML systems that govern much of our lives? The answer is no. There are algorithms, particularly in the field of the ML, that because of the mismatch between their own nature and the type of explanations that demand our understanding, are not intelligible no matter how much knowledge we possess on mathematics, computation, or any other related science.

3.3 Opacity as cognitive mismatch

As the last barrier of opacity, Burrell points out two difficulties: the size of large computer systems and the specific problems of ML algorithms (Burrell 2016). The first problem is not new. Joseph Weizenbaum already warned of it in the 1970s: some computer programs are so complex that none of the engineers involved in its development and maintenance can have a detailed understanding of the whole (Weizenbaum 1976). What engineers have to understand what other pieces of software do and that should fit with the ones they develop is summaries, reports; but the written word, as King Ammon warned Teuth in Phaedrus' Platonic myth about the invention of writing, entails the risk that those who use it not as a reinforcement for their memory, but to understand new things, believe that they understand what in reality they ignore (Phaedrus, 275a). That makes the large computer systems impenetrably opaque, with the consequent loss of control, "The systems of rules and criteria that are embodied in such computer systems become immune to change, because, in the absence of a detailed understanding of the inner workings of a computer system, any substantial modification of it is very likely to render the whole system inoperative and possibly unrestorable. Such computer systems can therefore only grow" (Weizenbaum 1976, 236).

Regarding the opacity of ML models, it is convenient to start by explaining how they work. As we said at the beginning, we are interested in the most opaque ones. But this category is still very wide (DARPA 2016, 5): support vector machines (SVM), random forests, probabilistic graphical models (PGM), reinforcement learning (RL), deep learning neural networks, etc. Here we will focus on artificial neural networks (ANN), which include deep neural networks (DNN), and we will do it for two reasons: the need to limit our scope and because these are the most opaque models and at the same time the most effective in classification and evaluation tasks.

Let's start by explaining what an ANN is with a simple example: an ANN for image recognition. An ANN is a set of units (neurons) organized in layers. The first layer, which is the one that receives information from the outside of the network, is called input layer. The last layer, which returns the result, is the output layer. In between there may be a variable number of hidden layers. The more the hidden layers, the greater the internal representation capacity of the network. Typically, each of the units of each layer activates each of the units of the next layer. This activation aims to emulate the electrical signals that travel in the nervous system of living beings from one neuron to the next. As in biological neurons, in artificial neurons there is a weight that determines the intensity with which one unit activates another. The product of each incoming activation by its weight is added, plus a number called bias multiplied by its weight, and the result is the input of an activation function that outputs the activation value of that unit over the next ones. In general, this is the structure and functioning of an ANN, although there are several variants.

For those who want a more precise insight, this is the formula that commonly rules the functioning of ANNs (Russell and Norvig 2016, 728):

$$a_j = f\left(\sum_{i=0}^{n} w_{i,j} a_i\right),$$

where a_i is the output activation for unit i, $w_{i,j}$ is the weight on the link from unit i to the current unit j, n is the number of units that activate a_j, a_0 is the bias activation, f is the activation function, and a_j is the output of the current unit j. The activation function f can be of many kinds: unit step (threshold), piecewise linear, gaussian, linear, etc. But the most common is the sigmoid function, because it allows to keep the output in reasonable numerical ranges and is non-linear, which is a key feature for solving complex problems.

ANNs usually include two operations or algorithms: a learning algorithm (learner) and a classification algorithm (classifier). For example, the aforementioned generalized delta rule was a breakthrough in the field of ANN learning made by David Rumelhart, James McClelland, and their team in the 1980s (Rumelhart et al. 1989). During the training, the network is provided a large number of samples and applies an algorithm such as the generalized delta rule (Rumelhart 1997). The objective of the process is to modify the matrix of weights, which in the beginning usually has random values, until it reaches a set of optimal values. The optimal values are those that produce

the most accurate possible classification of inputs. Burrell illustrates this with the case of an ANN trained to recognize digits from 0 to 9 written by hand in a bounded space of 8×8 pixels, that is, a total of 64 pixels, each of which stimulates a unit of the input layer with a variable intensity depending on the pixel brightness. After the training, the inputs provided to the network are no longer used to modify the matrix of weights, but rather the matrix of weights obtained from the training classifies the information inputs, returning through the output layer the digit corresponding to the handwriting that has been used as input.

The interesting thing here is that the network of the example performs the task of recognizing and classifying digits in a way that we could describe as irrational, as it is unintelligible. If its process were intelligible, says Burrell, it would proceed by decomposing the main task into other simpler subtasks intelligible to a human being, such as identifying a horizontal bar, a closed oval shape, a diagonal line, etc. (Burrell 2016). But it does not. It works similar to our visual cortex: in a way that even we cannot consciously explain. And when we try to do it, as it happens to patients with agnosia, our performance is clumsy. ANNs are especially good at this: solving problems for which we do not have an algorithm or a well-defined logical sequence of cognitive actions that produces the result we intend. ANNs are better at recognizing images than any symbolic program, that is, intelligible.

Although the simplest way to explain the functioning of an ANN is with an example of visual classification, the truth is that the potential of these machines goes much further. For example, for the evaluation of loan qualification through ANN, not a handful of pixels are used as inputs (features), but a multitude of relevant variables: age, sex, educational level, type of employment, historical annual income, marital status, number of children, criminal records, etc. To those obvious variables for their relevance in the determination of the loan qualification must be added others that are not so, but they are highly informative due to the correlations discovered by ML algorithms. A striking example is that of Deloitte executives who revealed in a conference that they can now "use thousands of 'non-traditional' third party data sources, such as consumer buying history, to predict a life insurance applicant's health status with an accuracy comparable to a medical exam" (Robinson et al. 2014). Thanks to this ML potential to discover correlations between phenomena as causally distant as the consumer buying history and the health status, legal limitations like those of the GDPR to avoid discrimination in automatic data processing are useless (Goodman and Flaxman 2016), since a data as apparently neutral as the postal code can, in combination with others, serve to deduce the race and include it as data in an automated decision process indirectly, without violating the article that prohibits discrimination based on race, sex, religion, etc.

In the age of universal access to the Internet, the trace we leave when surfing the net crystallizes in thousands of apparently insignificant data that, when supplied to a large ML system, can return as output an extremely precise profile of who we are. And how does the machine do that? In a way as opaque as the digit recognition process described, this is without words, without sentences, without arguments. Modern ANNs understand the world intuitively, as Nietzsche liked to say, which is the most effective kind of thinking, the one that does not require reason. Observing the huge matrix of weights of a DNN and the activation levels of the units in real time, Ortega y Gasset would say, may give an esthetic impression, but it does not explain how the computer reaches its conclusions.

4 Techniques to remedy opacity

Opacity as a cognitive mismatch in ML is so worrying that Defense Advanced Research Projects Agency (DARPA), a U.S. military technological research agency linked to AI since its inception (Nilsson 2009), launched the XAI program (Explainable Artificial Intelligence) in May 2017, with the goal of creating "a suit of new or modified machine learning techniques that produce explainable models" (DARPA 2016, 5). The XAI program, which will conclude in 2021, completed the first phase at the end of 2018, but as typical of military affairs, the results have not been made public. So we cannot account for some of the best funded advances in the race for shedding light on ML models. However, within the public domain there already are very interesting operational proposals. However, prior to examining these proposals it is convenient to fix a couple of definitions.

The opposite concept to opacity is usually interpretability. A model is interpretable when at least one interpretation can be made of it. This is a good definition of interpretation: "An interpretation is the mapping of an abstract concept (e.g. a predicted class) into a domain that the human can make sense of" (Montavon et al. 2017, 2). The authors use images (arrays of pixels) and texts (sequences of words) as examples of interpretable domains, as opposed to non-interpretable domains such as abstract vector spaces or sequences with unknown words or symbols. A second important definition is that of explanation, "An explanation is the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression)" (Montavon et al. 2017). The authors give as an example the heatmap of the input image that highlights the pixels that have contributed most strongly to the decision on the classification. In the example of the ANN that classifies digits, an explanation would be a heatmap, and a

heatmap would be a version of the input image to be classified in which the most decisive features deciding that the number is this one have been highlighted.

Academic research work done in recent years to remedy the opacity as cognitive mismatch improving the interpretability and the explainability of ML models is abundant (Bach et al. 2015; Nguyen et al. 2016; Ribeiro et al. 2016; Simonyan et al. 2014; Zeiler and Fergus 2013). The purpose of this section is to describe and evaluate all these within the limitations of space and topic (ANN). To describe them, it is convenient to primarily group them in categories. A fairly good classification is that of Zachary Lipton, who divides them into two large groups, depending on whether they satisfy the properties of transparency and post hoc interpretability (Lipton 2017). As for the evaluation, it consists of determining the degree to which they satisfy their objective. DARPA believes that the goal of explanation techniques is threefold: at the most basic level, to explain the individual decisions of a model; at an advanced level, to explain the strengths and weaknesses of the overall model; and at the most sophisticated level, enable the user to identify and correct mistakes (DARPA 2016, 13).

A technique introduces transparency when it allows to answer the question "how does the model work?". Instead, it introduces post hoc interpretability when it answers the question "what else can the model tell me?" (Lipton 2017). Lipton distinguishes three types of transparency: (1) at the level of the whole model (simulability) when its computational operations are simulable by a human being in a reasonable time, (2) at the level of the individual components (decomposability) when each part of the model (inputs, parameters, calculations) admits an intuitive explanation, and (3) at the level of the training algorithm (algorithmic transparency). Simulability limits the size and complexity of computations, decomposability limits the level of abstraction, and algorithmic transparency is simply not a property of opaque ML models, particularly ANNs. Some linear models have algorithmic transparency, but ANNs perform non-linear transformations.

As for post hoc interpretability, it is the most interesting from the technical point of view, because that is where the greatest progress has been made. Lipton divides post hoc interpretability techniques into four types: (1) text explanations, (2) visualization, (3) local explanations, and (4) explanation by examples. A model offers text explanations (1) if it verbally justifies its decisions. One way to do this is by training a second model to generate explanations for the first one (Krening et al. 2016). In the example of the model of ML that decides the loan qualification of a subject, after having trained it, a second ML model would have to be trained, taking as inputs the inputs and outputs of the first, and offering as outputs the verbal justifications given by human experts about why the first has granted or denied a loan in each of the training cases. This way, the most basic objective of explanation techniques would be met, that is, explaining individual decisions, but at the expense of introducing a second model that raises the total degree of opacity of the system, since there is a second model whose behavior must be explained. Introducing a third to explain the second would lead to an ad infinitum.

Regarding the visualization techniques (2), they are the most developed. An example would be the aforementioned heatmaps. There are several ways to generate heatmaps. Two very common are sensitivity analysis (SA) and layer-wise relevance propagation (LRP) (Samek et al. 2017). SA returns a heatmap on the assumption that the most relevant input features are those to which the output is more sensitive, while LRP redistributes the final prediction backwards, assigning more relevance to the most activated units and to those that are connected by stronger weights. Both can be applied to state-of-the-art DNNs, such as GoogleNet, as well as to SVM; both ML systems are very common. Activation maximization (AM) is also relevant, as it allows to visualize the prototype or the multiple local prototypes contained in a DNN through the search of the input values that would produce a maximum response of the model (Montavon et al. 2017). The visualization techniques, in short, allow to explain individual decisions.

However, from the point of view of Logic, indicating which inputs have been the most relevant to produce an output is not an explanation. And it is not for three main reasons. First, because it would be necessary for the inputs to have the aforementioned property of decomposability, that is, to be intuitively interpretable, which occurs when the inputs are images, but not when they are, for instance, highly engineered data. Second, because inputs are only a subset of the premises: the rest remain hidden in the form of internal representations of the model coded in the matrix of weights, and may vary when the training is extended. And third, because the sequence of inferences that enable to pass from the premises to the conclusion is not indicated. To believe that a heatmap, whether of images or texts, is an explanation, is to incur in a *fallacia no causae ut causae*; unless of course, we redefine what an explanation is in the terms previously seen, that is, understanding the explanation in a non-nomological-deductive sense, but systematic. An explanation is systematic when it is morphological (it describes a specified structure and the specific abilities of whatever is so structured) with the additional element of organized cooperative interaction (Haugeland 1981, 247). For example, the car's engine mentioned by Ortega y Gasset is a typical phenomenon suitable for a systematic explanation. Our reply is that if a heatmap is intended to be a systematic explanation, then it must meet the first two requirements mentioned: decomposability and exposure of the internal representations of

the model. But the latter is precisely the most opaque element of ANN models.

Third, local explanations (3) are an adequate technique for ANNs too complex to visualize the totality of what has been learned. A common way to do it in DNNs is by computing saliency maps. The disadvantage of local explanations is that they do not account for the whole model, but only for one part, so they tend to yield misleading results.

And finally, the explanations by examples (4) is a technique that, as the name suggests consist in that the model, in addition to returning an output, indicates cases similar to the one analyzed to persuade the user that the decision is correct. As Lipton points out, it is the equivalent to the human arguments from analogy. That arguments from analogy are weak is something well known in Logic (Douglas 2008). Therefore, explanations by example allow explaining individual decisions, but not conclusively.

5 Conclusion

Ultimately as of 2019, the words written by Nick Bostrom in 2014 in his popular book *Superintelligence* remain true: neuromorphic AI (ANN) should be avoided to not to worsen the problem of control over computers (Bostrom 2014, 301). ANNs are the most effective ML architecture, but also the most opaque. It is true that in the last few years great advances have been made to make the most opaque ML models interpretable and explicable, and it is expected that this field will continue to advance thanks to regulations such as the GDPR and to research efforts such as DARPA's XAI program. However, in the present, trusting in these models is as much as relying on boxes that, although they cannot be said to be black, are very dark.

Is it rational to trust classifications and evaluations made by agents that process information in an unintelligible way? Lipton points out that transparency is a design criterion that we demand from AI but not from other human beings (Lipton 2017). What he means is that we do not have access to other human beings' minds or brains, and we do not demand it to trust them. We only have access to their behavior and their explanations, which are not a reliable depiction of the real mental processes (Kahneman 2012). So Lipton is right, we ask the machines more than we ask our peers. However, it is a justified demand insofar as, although it is impossible to access the thoughts of other human beings, evolution has endowed us with mirror neurons and theory of mind to infer how they do it; tools that have proven to be effective over millennia (Iacoboni 2008). Instead, the computer is a mystery to us; a mystery that gets darker the farther it moves away from being a simple pocket calculator and the closer it gets to AI, be it symbolic or subsymbolic (Carabantes 2016; Benítez 2011; Copeland 1993). If the AI is symbolic, then it is hardly comprehensible by the user, because its heuristic rules, which act as our cognitive biases, are different and also tend to the minimum to explore the whole space of computationally possible solutions. And if the AI is subsymbolic, like ANNs, then it is incomprehensible even for its programmer because the operations that transform inputs into outputs are not compatible with human cognition—there are no words, no sentences, no arguments.

But, on the other hand, the smarter it is and the more it moves away from our understanding, the greater is its effectiveness: to predict if someone will repay a loan, or to foresee if someone will be profitable as beneficiary of a health insurance, or if someone is more likely to vote for this or that political party in the next election. The ANN predicts it all faster, cheaper and more accurately than human beings. And what company or government would give up such power of prediction? "Prediction is not just one of the things your brain does. It is the primary function of the neocortex, and the foundation of intelligence" (Hawkins and Blakeslee 2005, 60). ANNs imitate the functionality of the neocortex, they learn from reality and use that information to predict the future given a fragment of information. In pragmatist terms this is intelligence, this is an ANN, and this is techno-science: to know to predict, to predict to act (*science, d'où prévoyance; prévoyance, d'où action*); said the father of positivism (Comte 1989). Predicting is power. In the case of ANNs, that power is like the one conjured by the sorcerer's apprentice: the price to pay is the loss of control. Our society has decided to invoke dark powers—that is, unintelligible—to increase domination. In a certain sense, we return to the myth, to the attempt to control natural and social reality by invoking forces that we do not understand. The myth is already enlightenment, and the enlightenment, in its highest level of development, the current one, has fallen into the myth in a more literal way than Horkheimer and Adorno foresaw (Horkheimer and Adorno 2002).

Is this rational? It depends on how rationality is understood. In terms of instrumental or subjective rationality, yes; the proof is that those who decide on the use of these machines have chosen to use them to maximize their profits. In terms of axiological or objective rationality, no. But we are not so naive as to think that, neither through journalism nor through school, our society will assume the idea that "an intelligent man is not one who can merely reason correctly, but whose mind is open to perceiving objective contents" (Horkheimer 2004, 38). Nowadays, this intelligent observation about the categorical imperative is still true: "The citizen who renounced a profit out of the Kantian motive of respect for the mere form of the law would not be enlightened but superstitious—a fool" (Horkheimer and Adorno 2002, 67). Black-box AI is, in short, a faithful expression of our civilization.

References

Bach S, Binder A, Montavon G, Klauschen F, Müller KR, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10(7):1–46. https://doi.org/10.1371/journal.pone.0130140

Benítez A (2011) Fundamentos de inteligencia artificial. Libro segundo: Inteligencia artificial clásica. Escolar y Mayo, Madrid

Benítez A (2013) Fundamentos de inteligencia artificial. Libro tercero: Inteligencia artificial bioinspirada. Escolar y Mayo, Madrid

Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press, Oxford

Brennan J (2016) Against democracy. Princeton University Press, Princeton

Bueno G (2006) Zapatero y el pensamiento Alicia: un presidente en el país de las maravillas. Temas de Hoy, Madrid

Burrell J (2016) How the machine 'thinks': understanding opacity in machine learning algorithms. Big Data Soc. https://doi.org/10.1177/2053951715622512

Carabantes M (2016) Inteligencia artificial: una perspectiva filosófica. Escolar y Mayo, Madrid

CFTC & SEC (Commodity Futures Trading Commission and Securities & Exchange Commission) (2010) Findings regarding the market events of May 6, 2010: report of the staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues. Washington, DC

Comte A (1989) Course de philosophie positive. Nathan, Paris

Copeland J (1993) Artificial intelligence: a philosophical introduction. Blackwell, Oxford

DARPA (Defense Advanced Research Projects Agency) (2016) Explainable Artificial Intelligence (XAI). DARPA-BAA-16-53. https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

De Bruin B (2015) Ethics and the global financial crisis. Cambridge University Press, Cambridge

De Bruin B, Floridi L (2016) The ethics of cloud computing. Sci Eng Ethics 23(1):21–39. https://doi.org/10.1007/s11948-016-9759-0

Diakopoulos N (2013) Algorithmic accountability reporting: on the investigation of black boxes. Report, Tow Center for Digital Journalism, Columbia University

Douglas W (2008) Informal logic: a pragmatic approach. Cambridge University Press, Cambridge

Dreyfus H (1992) What computers still can't do. The MIT Press, Cambridge

Ferrater J (2009) Diccionario de Filosofía. Ariel, Barcelona

Fraunhofer (2017) Watching computers think. Fraunhofer Research News (blog). http://www.fraunhofer.de. Accessed 7 Dec 2018

Goodman B, Flaxman S (2016) European Union regulations on algorithmic decision-making and a "right to explanation". AI Mag 38(3):50–57. https://doi.org/10.1609/aimag.v38i3.2741

Haugeland J (1981) The nature and plausibility of cognitivism. In: Haugeland J (ed) Mind design. Cambridge University Press, Cambridge, pp 243–281

Hawkins J, Blakeslee S (2005) On intelligence. Times Books, New York

Horkheimer M (2004) Eclipse of reason. Continuum, London

Horkheimer M, Adorno T (2002) Dialectic of enlightenment. Stanford University Press, Stanford

Iacoboni M (2008) Mirroring people: the new science of how we connect with others. Farrar, Straus and Giroux, New York

Kahneman D (2012) Thinking, fast and slow. Penguin Books, New York

Krening S, Harrison B, Feigh K, Isbell C, Riedl M, Thomaz A (2016) Learning from explanations using sentiment and advice in RL. IEEE Trans Cogn Dev Syst 9(1):44–55. https://doi.org/10.1109/TCDS.2016.2628365

Lapuschkin S, Binder A, Montavon G, Müller KR, Samek W (2016) Analyzing classifiers: fisher vectors and deep neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2912–2920. https://doi.org/10.1109/CVPR.2016.318

Lipton Z (2017) The mythos of model interpretability. http://arxiv.org/abs/1606.03490v3. Accessed 26 Dec 2018

McCarthy J, Hayes P (1969) Some philosophical problems from the standpoint of artificial intelligence. In: Meltzer B, Michie D (eds) Machine intelligence, vol 4. Edinburgh University Press, Edinburgh

Montavon G, Samek W, Müller KR (2017) Methods for interpreting and understanding deep neural Networks. Digit Signal Process 73:1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Nguyen A, Dosovitskiy A, Yosinski J, Brox T, Clune J (2016) Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. http://arxiv.org/abs/1605.09304v5. Accessed 10 Dec 2018

Nilsson N (2009) The quest for artificial intelligence: a history of ideas and achievements. Cambridge University Press, New York

Ortega y Gasset J (2009) Introducción al curso ¿Qué es la técnica? In: Obras completas, vol IX. Taurus Ediciones, Madrid, pp 27–31

Pasquale F (2015) The black box society: the secret algorithms that control money and information. Harvard University Press, Cambridge

Ribeiro M, Singh S, Guestrin C (2016) Why should I trust you?: explaining the predictions of any classifier. http://arxiv.org/abs/1602.04938v3. Accessed 27 Dec 2018

Robinson D, Yu H, Rieke A (2014) Civil rights, big data and our algorithmic future. Social Justice and Technology. https://centerformediajustice.org/wp-content/uploads/2014/10/Civil-Rights_Big-Data_Our-Future.pdf. Accessed 4 Jan 2019

Rumelhart D (1997) The architecture of mind: a connectionist approach. In: Haugeland J (ed) Mind design II. Cambridge University Press, Cambridge, pp 205–232

Rumelhart D, McClelland J, The PDP Research Group (1989) Parallel distributed processing, vol I. The MIT Press, Cambridge

Russell S, Norvig P (2016) Artificial intelligence: a modern approach. Global Edition, 3rd edn. Pearson Education, London

Samek W, Wiegand T, Müller KR (2017) Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. http://arxiv.org/abs/1708.08296v1. Accessed 26 Dec 2018

Sandvig C, Hamilton K, Karahalios L, Langbort C (2014) Auditing algorithms: research methods for detecting discrimination on internet platforms. In: Annual Meeting of the International Communication Association, Seattle, WA, pp 1–23

Simonyan K, Vedaldi A, Zisserman A (2014) Deep inside convolutional networks: visualizing image classification models and saliency maps. http://arxiv.org/abs/1312.6034v2. Accessed 10 Dec 2018

Somin I (2013) Democracy and political ignorance: why smaller government is smarter. Stanford University Press, Stanford

Weizenbaum J (1976) Computer power and human reason: from judgement to calculation. W. H. Freeman & Company, San Francisco

Zeiler M, Fergus R (2013) Visualizing and understanding convolutional networks. http://arxiv.org/abs/1311.2901v3. Accessed 6 Dec 2018