ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ

Εθνικόν και Καποδιστριακόν
Πανεπιστήμιον Αθηνών

Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

Τμήμα Ιστορίας και Φιλοσοφίας της Επιστήμης
&
Τμήμα Πληροφορικής και Τηλεπικοινωνιών

Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών:
«Science, Technology, Society—Science and Technology Studies»

Διπλωματική Εργασία (MSc Thesis)
“The Black Box of AI»

Anestis Karastergiou

Αριθμός Μητρώου: 11/220

Τριμελής Συμβουλευτική Επιτροπή:
Ευστάθιος Ψύλλος, καθηγητής (επιβλέπον μέλος)
Αριστοτέλης Τύμπας, Καθηγητής (μέλος)
Εμμανουήλ Σίμος, Μεταδιδακτορικός Ακαδημαϊκός Υπότροφος (μέλος)

Αθήνα 2021

HELLENIC REPUBLIC
National and Kapodistrian
University of Athens

National and Kapodistrian University of Athens

Department of History and Philosophy of Science
&
Department of Informatics and Telecommunications

Interdepartmental Graduate Program:

Science, Technology, Society—Science and Technology Studies

MSc Thesis

“The Black Box of AI”

Anestis Karastergiou

Registration Number: 11/220

Thesis Advisory Committee:
Stathis Psilos, Professor (advisor)
Aristotle Tympas, Professor (member)
Emmanouil Simos, Postdoctoral Fellow (member)

Athens 2021

Abstract

In this study, AI is critically analyzed through the lens of the black box metaphor, a useful methodological tool within the field of science and technology studies. Beginning with an overview of the secondary literature and some philosophical issues related to AI and the black box concept, this thesis proceeds to an analysis of primary data taken from the scientific journals *Nature* and *Scientific American*. The main purpose of this work is to show how the black box of AI is constructed, to present the nuances related to it, how to unpack it and, in general, to present the full extent to which one can use the black box metaphor in order to reflect on AI. Attempting to answer basic research questions about AI's black box, its construction through a co-production of technoscientific and social factors is revealed along with ways to pry open its black box. Understanding the construction of the black box is the first step in order to start opening it. Its implications for society lead to the mandate for transparency in order to mitigate the adverse effects stemming from blackboxing procedures. Contingent on the black box of AI and the transparency mandate is trust in AI. Increasing trust in AI presents a difficult but necessary task as AI is, ultimately, a technosocial phenomenon that both shapes and is shaped by society. Finally, in this thesis the relationship between the biological and the mechanical domain by means of the black box metaphor and AI research is evaluated.

Table of contents

1. Introduction

The purpose of this thesis is to present the black box of artificial intelligence (AI) and the nuances thereof. Starting with an overview of the literature regarding AI, the proliferation of books and articles on the subject is approached mainly through the lens of science and technology studies (STS). A philosophical definition of AI is mentioned in order to set the appropriate boundaries within which the discussion should take place. A brief historical analysis of AI and its main constituents does not attempt to describe an accurate periodicity of it, as this is a controversial issue within the circles of the history of science. It attempts to set the schema in order to see how the discussion around AI has progressed and to set the frame for understanding its current boom. In other words, the first part of this thesis, the secondary literature review regarding AI and its black box, does not present an extensive review of past and current literature but an overview that provides the tools to process writings on AI.

Key issues regarding the "black box" concept are also presented in the first part through an analysis of STS literature regarding the black box metaphor. Beginning with the roots of the metaphor, one proceeds to understand its utilization as a methodological tool to analyze AI concepts, such as big data, machine and deep learning. The second part of the thesis covers the presentation and analysis of primary data. More specifically, articles on the theme of AI and its black box are collected from the journals *Nature* and *Scientific American*. Through a textual analysis of these, one can interpret how the co-production of AI's black box is depicted, how the methods of its unpacking are perceived and what are the implications of the black box metaphor for society and technoscience. After presenting these articles, an interpretation of them through an STS prism follows.

In reference to machine blackboxing, Latour has pointed out that "…the more science and technology succeed, the more opaque and obscure they become" (Latour, 1999, p. 304). Following a review of relevant historical, philosophical and STS literature on the black box of AI, we will endeavor to apply Latour's thesis on AI machines. Undoubtedly, deep learning instances are very complex and opaque. Modeled, however loosely, on the brain, the archetypal black box, they present the pinnacle of technoscientific progress in AI. Questioning the idea—i.e., the linearity—of progress may not be appropriate here. Suffice it to say that the idea of linear progress in technoscience is appealing but very problematic. For example, computational power increases exponentially, but, on the one hand, its limits seem to be within sight already, requiring different ways to solve problems apart from brute force (Pacheco, 2011; Hwang, 2018). On the other hand, simply mentioning quantifiable measures of progress does not actually exhaust the definition of the term. To be precise, when Latour talks about success, he appears to have the idea of progress in mind. Scientific progress is measured by its successes (Latour, 1999, p. 304).

Consider the case of a deep learning algorithm used to predict the epidemiological load of the current SARS-Cov-2 pandemic. If the results are correct, i.e., the answers the machine gives seem to help achieve our end result, which is none other than holding the diffusion of the virus at bay, the implementation of such an algorithm is successful. If the results were wrong or perceived as such, corrective actions would take place, such as finding and analyzing what went wrong or even abandoning its use. When a technoscientific device is perceived as successful, no one feels the need to inquire about the processes involved in the production of results. Its creators are content, along with policy makers and, eventually, society. It seems that perceived success leads to the co-production of the algorithmic black box. The more successful it is perceived to be, the more robust its black box becomes. This is exactly what Latour means when he says that technoscientific success makes the black box opaquer. On the contrary, failure or perceived failure brings to light the need to pry open the machine's black box.

However, as we tried to point out, success and progress in science are controversial concepts and, largely, a matter of perception. They are co-produced by the interplay of technoscience, society and politics as an attempt to assign positive value to scientific practice. But perceived success or even "actual" success does not guarantee future success, and progress should not be seen as a straight ascending line towards better results. A "whatever works" attitude may seem plausible in many cases, but it is bound to create an unequal distribution of perceived successes by generalizing from previous ones while turning a blind eye on the social and ethical implications of technoscientific "success". Explaining scientific process, thus fostering transparency, could be a first step to open the black box instead of letting it take over technoscience and society. Having this critique in mind, one can now understand why elucidating the black box problem, critically reflecting on it and, finally, attempting to resolve it by opening the black box is of paramount importance in the digital era.

2.Introduction to Secondary Literature and Philosophical Issues Concerning AI

2.1 Literature on AI

In order to provide a comprehensive review on AI, we will attempt to set the appropriate framework within which we propose to analyze the vast amount of literature existing today. It starts with a philosophical definition that points to the actual implementation of AI nowadays. There is probably too much attention drawn to the prospect of an AI that could equal or even exceed human intelligence. However, this type of AI, i.e., strong or general AI, is currently science fiction material and is bound to remain so in the near future, as will become evident from the discussion that follows. Even if modern computing capacity is exceeding expectations (Hao, 2019) and big data is growing at unprecedented rates, it seems that there is a lot more to be done before reaching strong AI.

Scientists have long fantasized about reaching the point of singularity, when AI becomes virtually indistinguishable from human intelligence. One instance of this science fiction can be seen in the TV series *Star Trek: The Next Generation*, where Data, a humanlike android that is fully capable of performing human tasks while outperforming humans in computational tasks, joins the crew. However, even in this case Data struggles with issues that are supposed to be uniquely human, such as emotions or, more significantly for our purposes, conscience (Hanley, 1997). To define conscience philosophically is an unfathomably difficult task and, of course, a pervasive problem that would take too much space even to schematically present it. However, it is important to note that strong AI should have to replicate conscience if it were to reach the point of singularity.

Apart from the existential issues raised in Data's case, there is another science fiction character that can bring us closer to modern AI and the black box issues related to it. This is none other than HAL 9000 from Arthur Clark's *2001: A Space Odyssey*, adapted for the big screen by Stanley Kubrick. HAL (Heuristically programmed Algorithmic Computer) is the epitome of AI as it is capable of performing virtually any task given to it, along with understanding and expressing human emotions. Again, the sensing and expression of emotions is what makes AI actually intelligent in the eyes of these pioneering thinkers. However, this is not the point here. HAL represents an AI that is fully capable of making decisions that profoundly affect humans. For instance, it is even responsible for the spaceship's life support systems. HAL's capabilities, which far exceed modern AI applications, bring fear to its human counterparts and the whole movie reaches its climax with the conflict between the two antagonists. From our point of view, two basic issues arise from HAL's depiction here, which are closely related. On the one hand, it involves the role of AI machines in society and what tasks would people be willing to let the machine perform (Raymond et al., 2017, pp. 250-254). On the other hand, HAL's black box and the difficulty to pry it open instills fear and distrust to AI. Many of HAL's capabilities would characterize it as a very advanced weak AI (such as advanced computer vision and the like). The prospect of such an AI machine seems much more plausible than reaching singularity, and this is why it can be used as a popular science fiction case to demonstrate societal issues regarding AI's black box.

In STS literature, the critique of the "smartness mandate" has gained momentum. The reason for its inclusion here has to do with its inextricable link to AI. Loosely put, the smartness mandate refers to the algorithmic transformation of society. In order to find a solution to the problems societies face today (environmental, economic and so on), being smart promises solutions built on resilience and the constant optimization of society. As stated by its critics:

> [I]nsofar as smartness separates critique from conscious, collective, human reflection—that is, insofar as smartness seeks to steer communities algorithmically, in registers operating below consciousness and human discourse—critiquing smartness will in part be a matter of excavating and rethinking each of its central concepts and practices (zones, populations, optimization, and resilience), as well as the temporal logic that emerges from the particular way in which smartness combines these concepts and practices. (Halpern et al., 2017, p. 125)

In other words, the mandate to be smart and to use AI for the optimization of society calls for a broader critique, which should include many relevant actors and not just rely on the partial truths derived from the promises of smartness.

2.1.1 AI from a Philosophical Perspective

Defining AI is a notoriously difficult task, in which we will delve into briefly and with caution. Starting with John Searle's distinction between strong and weak AI, we will be able to set the appropriate frame in which to talk about AI and data ethics. According to the philosopher, for "weak AI, the principal value of the computer in the study of the mind is that it gives us a very powerful tool" (Searle, 1980, p. 417). He goes on to mention that "for example, it enables us to formulate and test hypotheses in a more rigorous and precise fashion" (Searle, 1980, p. 417). In other words, AI is a tool we have in our hands and a very powerful one for that matter. We construct the algorithms and the neural networks, we feed them data and they provide us with outputs, which can vary according to its abilities to learn and adapt to new problems. It is quite obvious, that this type of AI is closer to machine and deep learning instances. However, strong AI is on a whole separate level. As Searle puts it:

> [A]ccording to strong AI, the computer is not merely a tool in the study of the mind; rather, the appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states. (Searle, 1980, p. 417)

He continues: "in strong AI, because the programmed computer has cognitive states, the programs are not mere tools that enable us to test psychological explanations; rather, the programs are themselves the explanations" (Searle, 1980, p. 417). To rephrase, the states of strong AI equal the "mental" states of the mind—i.e., strong AI equals conscience. Of course, conscience is a huge philosophical issue, and there is no room in this thesis for such a digression. However, the main point here is that strong AI is unattainable up to now, and it seems that it will remain so for the near future, to say the least. However, weak AI already exists, it has many applications in our daily lives, and it is bound to expand in the near future. Also, it is important to mention that weak AI applications and machine/deep learning instances need data; big data. This is why a separate section is devoted to Big Data, the building block of modern AI.

From a philosophical perspective, approaching AI solely through Searle's distinction between weak and strong AI, we would like to stress and problematize his assertion that in strong AI the machine states are themselves the explanations. Given that the machine states equal mental states, his conclusion seems perfectly plausible in the domain of philosophy of mind and identity theory (Smart, 2017). A mental state cannot and, probably, should not be explained by virtue of another one. The mental state is in itself the psychological explanation. Undoubtedly, it is a huge leap and by no means intended by the philosopher, but what if we link this statement to the black box metaphor within the STS framework? What are the repercussions of such a definition of strong AI for AI's black box? Simply and somewhat naively put, it seems that the opaquer the machine state becomes, or alternatively the harder its black box becomes, the closer we get to strong AI. Blurring again the boundaries between philosophy and STS by means of the black box metaphor, consider Nagel's article "What Is It Like to Be a Bat?" (Nagel, 1974). Setting aside the main purpose of the article, which is to discuss the mind-body problem, consider the limitations of physicalism in the explanation of mental states and raising the problem of objective and subjective experience, among others, the mental state of the bat, as any mental state, is in itself a black box. Its subjective nature prevents it from being explainable. At least up to now, it is only self-explanatory. Projecting this view of mental states, i.e., the black box that explains itself by virtue of itself, to machine states is quite problematic from our point of view and should not be a criterion for strong AI. For instance, suppose there is a deep learning machine state so complex that it defies explanation. We are incognizant of how it was reached, and we do not have the means to analytically represent it. The easy way out would be to consider it self-explanatory. This way, we ascribe subjectivity to the machine and, given that subjectivity is closely linked to human intelligence, we think we have reached singularity. However, it is not certain that our current lack of means to pry open the machine's black box, as the mental state one, will be the case in perpetuity. Here we argue that black boxes are to be cracked however hard they might appear and that reaching strong AI should not be through a process of elimination. In this light, we consider any AI black box, even deep learning ones that seem too hard to crack, within the boundaries of weak AI.

2.1.2 AI History: A Brief Overview

Outlined below, there are several definitions that give a fairly coherent picture of the basic features of AI. One of them is the definition given by John McCarthy, according to which AI is the science of "making intelligent machines, especially intelligent computer programs" (McCarthy, 2007, p. 2). However, he adds that artificial intelligence does not have to be limited to methods that are "biologically observable" (McCarthy, 2007, p. 2). Another definition is that of the well-known researcher in the field of AI, Patrick Henry Winston. According to him, AI lies in the ability of computers or robots to perform tasks that humans normally do (Winston, 1993, pp. 5-12). Its meaning may also include the development of computer systems that perform intelligent processes. In other words, machines perform reasonable tasks intelligently (Winston, 1993, pp. 5-12). At this point, it is worth mentioning that AI can be divided into strong and weak or limited AI. The first does not even exist yet and may be exceedingly ambitious as its goal is to reach human intelligence. On the other hand, limited AI is the form of AI that prevails today.

Through this brief list of definitions there seem to be some common ground as well as differences. For this reason, a broader definition would be better. In particular, McCarthy defines intelligence as the computational part of reasoning that is oriented towards achieving certain goals (McCarthy, 2007, p. 9). There are admittedly a variety of types and degrees of intelligence that exist in both humans and animals, and even in some machines. However, he believes that there is no definition of intelligence that is not based on a comparison with human intelligence. Similarly, Arthur R. Jensen, a leading researcher in human intelligence, proposes as a working hypothesis that all normal people have the same mental mechanisms and that differences in intelligence lie in quantitative, biochemical and physiological circumstances (Jensen, 1980, pp. 103-110). Jensen detects them in speed, short-term memory and the ability to form accurate and recoverable long-term memories (Jensen, 1980, pp. 103-110; Rushton & Jensen, 2010, pp. 9-12).

Even if Jensen is right about human intelligence, what we learn from AI points to the opposite conclusion. Computer programs have high speeds as well as a large amount of memory. However, their capabilities are limited to the mental mechanisms that the designers of these programs understand well enough to translate into a program (McCarthy, 2007, p. 4-5). According to Winston, the issue is further complicated by the fact that the cognitive sciences have not yet been able to determine exactly what human mental abilities are (Winston, 1993, pp. 33-37). It is possible that the organization of mental mechanisms in AI has some utility different from that of human intelligence. However, when developers try to emulate human intelligence and fail, it shows that they cannot understand it well enough yet. Winston believes that artificial intelligence rarely tries to simulate human intelligence, which is indeed the case with the limited AI mentioned above (Winston, 1993, p. 8). AI researchers are free to use methods that are not observed in humans or that involve far more calculations than humans can perform (Winston, 1993, pp. 8-12). All of the above show that the term AI is very broad, and the field of AI should be approached as open-ended. However, for the purposes of this thesis it is more interesting to present a brief history of this area.

The conventional beginning of AI can be found in the Dartmouth working group conference that took place in 1956. The organizers, John McCarthy and Marvin Minsky, drafted their proposal for a program in AI that would involve ten researchers and would last about two months (McCarthy, 1955, pp. 1-5). There was presented what is considered by many as the first AI program, Herbert Simon's "Logic Theorist" (Russel & Norvig, 1995, p. 17). However, the origins of the idea of artificial intelligence can be traced back to Alan Turing's landmark article "Computing machinery and intelligence" from 1950. In this article, this leading mathematician first posed the question of whether and in what terms machines can think, while he also set the conditions for his answer (Turing, 1950). The famous Turing test is a key criterion for concluding whether or not a machine has intelligence. To date, attempts by a machine to pass this test have either failed or yielded controversial results. Of course, the test has changed considerably since its original formulation, but the theoretical background remains more or less the same. However, if someone wants to speak according to the terms already mentioned, this test is an obstacle only for strong AI.

Regardless of the Dartmouth working group's achievements, one cannot downplay the contribution of this event to the history of AI. Later on, from 1957 to 1974, AI flourished. Mechanical algorithms have evolved considerably. The contribution of programs such as Newell and Simon's General Problem Solver to problem solving and Joseph Weizenbaum's ELIZA (1923-2008) to natural language processing has been remarkable (Weizenbaum, 1966, p. 36). These developments have sparked interest from various sponsors, including government agencies such as the Defense Advanced Research Projects Agency (DARPA) in the United States. With a slight pause, AI flourished again from 1980 onwards, as algorithms and sponsorship opportunities expanded. Computers such as Deep Blue, which defeated Gary Kasparov in chess in 1997 (Pandolfini, 1997, pp. 65-66), and Alpha Go, which defeated in the Chinese Go the professional Ke Jie, have begun to show how fast the evolution of artificial intelligence can go (Chen, 2016, para. 1). The main reason why these computers were victorious is not so much their ingenious algorithms as their ability to accumulate huge amounts of data and process them with, admittedly, astonishing speed. The question, then, is how these computers managed to accumulate so much information. The answer is, of course, simple. It is provided by the human agent.

2.1.3 Big Data

To begin with, big data can be defined, according to SAS[1], as:

> [A] term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it's not the amount of data that's important. It's what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves. ("Big Data: What it is and why it matters?", 2021, para. 1)

In addition, Oracle's[2] definition is as follows:

> [B]ig data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can't manage them. But these massive volumes of data can be used to address business problems you wouldn't have been able to tackle before. ("What Is Big Data?", 2021, para. 1)

One uptake from these definitions is, of course, that big data consists of huge amounts of data that can be collected and analyzed in unprecedented levels using modern technology. The importance of big data in AI applications can be stressed more by recalling the words of Peter Norvig, one of Google's directors, who said "We don't have better algorithms. We just have more data" (Sanders, 2014, pp. 3-18). In other words, data expansion is crucial to machine and deep learning.

[1] SAS is a leading organization in analytics, artificial intelligence and data management. For reference: https://www.sas.com/en_us/home.html.

[2] Oracle is an American multinational computer technology corporation. For reference: https://www.oracle.com/index.html.

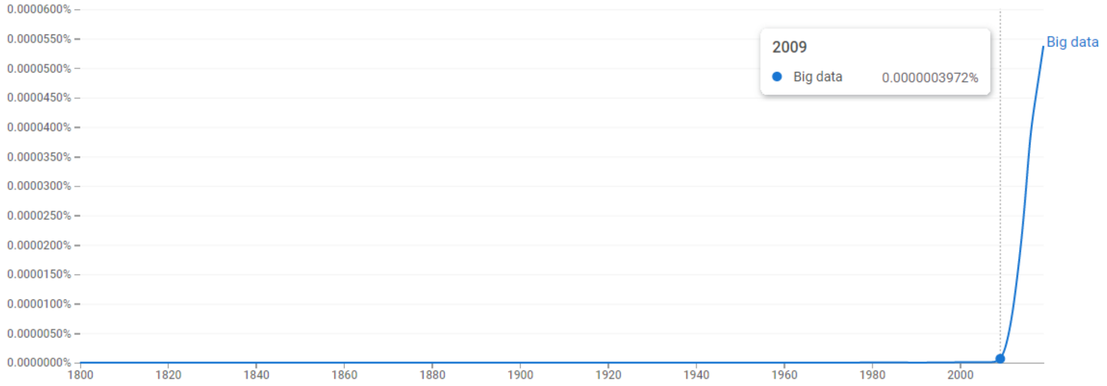

Figure 1: Google Ngram for the term “big data”.

Following a brief discussion of big data, machine learning should follow in the schema of dividing AI into its constituents. However, as machine learning is an umbrella term that includes deep learning, we proceed to our discussion of the latter. After all, it can be seen as the most sophisticated form of machine learning, a technology capable of making AI’s black box opaquer and, of course, the main target of our analysis.

2.1.4 Deep Learning

> What's in a name? That which we call a rose
> By any other name would smell as sweet
> William Shakespeare, *Romeo and Juliet*, ca. 1600

Nowadays, “deep learning” is a term that refers to cutting-edge machine learning technology. Geoffrey Hinton is considered to be the father of deep learning as he, along with some colleagues, presented a paper in 2006, titled “A Fast Learning Algorithm for Deep Belief Nets”. In this paper they articulated the basics for a fast-learning algorithm that can be used to create deep learning neural networks, which they named “Deep Belief Nets” (Hinton et al., 2006, p. 1527). Without going into the details of this algorithm, it is important to mention that it is akin to the stochastic gradient descent used in modern day deep neural networks in order to increase learning speed. More specifically, stochastic gradient descent is used to optimize gradient descent as it samples a subset of summand functions, thus facilitating large-scale machine learning (Bottou, 2010, pp. 177-183). Therefore, instead of descending from each point of the function, which would imply a huge number of iterations for big data, using samples stochastic gradient descent minimizes the number of iterations and is relatively unaffected by any changes in the amount of data (Bottou, 2010, pp. 177-184). However, the birth of the term “deep learning” should be traced further back to 1986, when Rina

Dechter introduced the term to machine learning (Dechter, 1986, 178-182). Moreover, in 2000, Aizenberg and his colleagues introduced the term to Artificial Neural Networks or ANNs (Aizenberg et al., 2000, pp. 9-24). One could now be tricked into believing that deep learning is something entirely new, a 21st century invention with roots that go back to 1986. The Ngram below can make this argument even stronger as one can see that the usage of the term begins around 2009 and then it spikes.

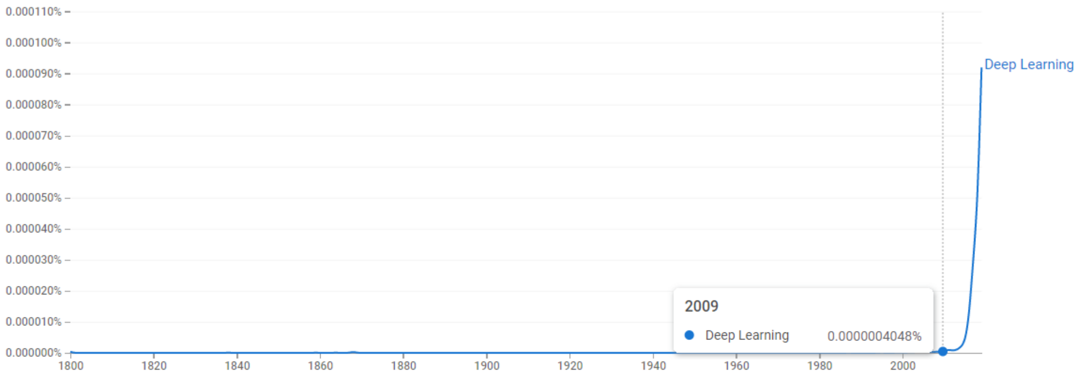

Figure 2: Google Ngram for the term "deep learning".

However, we will illustrate below that this is actually not the case. The idea along with the model can be traced further back. After this genealogy of deep learning, we will proceed to an analysis of the black boxing procedures related to it and how to unpack them, should such a thing be possible.

One has to go back to the 1940s in order to trace the origins of deep learning, i.e., neural networks. This is the decade of cybernetics as it will be exemplified in Norbert Weiner's book *Cybernetics: Or Control and Communication in the Animal and the Machine*, which came out in 1948 to establish the term (Wiener, 2000). However, the work of McCulloch and Pitts back in 1943 is crucial for our exercise here. In their paper, neural activity was treated for the first time as something that can be analyzed with the application of propositional logic. Undoubtedly working within the frame of logical positivism, they created the foundations for a linear algorithm, a simple neural network (Nilsson, 2010, pp. 34-43). Later, in 1950, Rosenblatt would coin the term "perceptron", a slightly more complicated algorithm, a neural network that can assign (that is, learn) the weights automatically. Needless to say that perceptrons form the basis of modern-day neural networks and, of course, deep learning. A multi-layer perceptron is probably the most common application in deep learning. Using logistical regression, back propagation and stochastic gradient descent, to oversimplify a complex procedure, deep learning neural networks can achieve fascinating results. However, their roots are reaching back to the 1940s and 1950s, the beginning of AI (Nilsson, 2010, pp. 92-102).

However, the early neural networks of McCulloch and Pitts did not actually learn. It is after Rosenblatt and some shallow attempts at supervised and unsupervised learning that we reach the predecessors of deep learning, i.e., the first feedforward multilayer perceptrons. Ivakhnenko and Lapa back in 1965 published the first general, working algorithm for supervised learning. In 1971, Ivakhnenko's deep network included eight layers trained by the 'Group Method of Data Handling' (Ivakhnenko, 1971, pp. 364-370), which remains popular in the new millennium since it learned to create representations of incoming data—for example, syntactic pattern recognition methods (Fukushima, 1979, pp. 658-663). Another milestone in training deep learning models that should be mentioned is Seppo Linnainmaa's 1970 master's thesis (Linnainmaa, 1970), which included a FORTRAN code for back propagation (the use of errors in training deep learning models), though this was not applied in neural networks until 1985. This was when Rumelhart, Williams, and Hinton demonstrated back propagation in a neural network could provide "interesting" distribution representations (Rumelhart et al., 1986, pp. 318-332). Nonetheless, following its promising birth, AI supposedly experienced its first winter around 1974 as it appeared impossible to live up to its promises. This could be also seen as the end of cybernetics, but that is a separate discussion. Winter came but did not last long. In 1979, Kunikho Fukushima developed an artificial neural network, called Neocognitron, with multiple pooling and convolutional layers, which used a hierarchical, multilayered design. Fukushima's design allowed the model to recognize visual patterns as well as important features to be adjusted manually in order to increase the 'weight' of certain connections (Fukushima, 1979, pp. 658-662).

The years between 1980 and 1987 are known as the period of the classicists. It was back then that artificial neural networks were actually introduced, and many modern-day techniques were applied, such as distributed representation and feedback, back propagation, and long short-term memory (LSTM). The latter is especially significant for deep learning as it uses feedback connections instead of just feedforward ones. It is considered to be the spearhead of contemporary deep learning. Obviously, its roots can be traced 40 years back in time. However dubious this historical periodization might be, an AI boom is followed by an AI winter and that came during the 1990s. But, at the end of this decade and the beginningof the 21st century AI has seen explosive growth (Nilsson, 2010, pp. 507-531, 656-657).

2.1.5 Trust in AI

Empirical research on the three types of AI representations sets out specific criteria that can promote human trust. These include tangibility and transparency so that the users are able to perceive how AI works, making its basic rules of operation obvious (Glikson & Woolley, 2020, pp. 11-12). Also, they include reliability, i.e., showing a steady behavior over time, as well as immediacy of behavior and finally the characteristics of AI tasks, whether they are of technical nature or require social skills (Glikson & Woolley, 2020, pp. 11-12).

Beginning with robotic AI, the physical presence of a robot has a positive effect on human cognitive trust (Glikson & Woolley, 2020, pp. 14-16). The more human it is, the more it is believed by users that it will make ethical decisions, but this does not mean that it is considered to be smarter (Glikson & Woolley, 2020, pp. 14-16). Transparency can increase trust, but the empirical research on the specific subject is scarce. The research has been done on robots that operate in remote areas (Glikson & Woolley, 2020, pp. 14-18). There is a general positive correlation between the constant flow of information from the robot and trust (Glikson & Woolley, 2020, pp. 20-21). The more they know about its functions, the more people trust it. In matters of reliability, when the robot is considered to have high intelligence, humans tend to trust even a defective robot. Studies have shown that in high-risk situations, participants lost confidence in the advice of a robot that made a mistake (Glikson & Woolley, 2020, pp. 18-20). In particular, trust decreased more when the mistake was made in the beginning than in later stages of the interaction (Glikson & Woolley, 2020, pp. 18-20). In cases of interaction with low-reliability robots, there was an increase in trust even though the robot made persistent mistakes. Ultimately, reliability could play a less important role in human confidence than expected (Glikson & Woolley, 2020, pp. 18-20). Concerning task characteristics, trust is increased in technical tasks rather than in jobs that require social intelligence (Glikson & Woolley, 2020, pp. 20-22). Once again, the human behavior of the robot tends to increase user confidence. Responsiveness, adaptability and social behaviors are the main factors that increase trust in immediacy behaviors (Glikson & Woolley, 2020, pp. 22-24). The incorporation of higher levels of intelligence into machines allowed robots to react to human presence and speech, creating social robots that can serve a social role (Glikson & Woolley, 2020, pp. 42-44). It has been observed that humans prefer a robot that acts on its own to one that works only when asked (Glikson & Woolley, 2020, pp. 42-44). Also, when human movements and gestures are developed, this has a positive impact on human perception of a robot's anthropomorphism (Glikson & Woolley, 2020, p. 53).

In virtual AI where there is no physical presence of the machine, the research focuses mainly on chatbots or avatars. In general, unlike robotic AI, the confidence trajectory suggests that high initial confidence decreases after each interaction. However, there are some indications of a relatively low initial confidence that increases upon interaction (Glikson & Woolley, 2020, pp. 24-26). Concerning tangibility, visualization and anthropomorphism in a virtual agent who has been given high intelligence can increase the levels of trust (Glikson & Woolley, 2020, pp. 26-27). Transparency can contribute in building trust by explaining to users how the system they interact with works and why the specific algorithm is used, but only when they are informed from the beginning about its level of reliability (Glikson & Woolley, 2020, pp. 27-28). On the other hand, low reliability mainly reduces trust in laboratory studies where initial confidence was very high (Glikson & Woolley, 2020, pp. 28-29). The results of the research showed that direct experience greatly reduces trust, inconsistency in reliability reduces trust more than low reliability, while the research tends to focus more on synchronization between user expectations and AI capabilities (Glikson & Woolley, 2020, pp. 28-29). When focusing on the characteristics of the tasks in virtual AI, it is noted that in technical tasks that require data analysis, trust in AI is even higher than in humans (Glikson & Woolley, 2020, pp. 29-30). When examining the immediacy of behavior, personalization tactics (such as personal questions to the user and the use of persuasion) increase trust (Glikson & Woolley, 2020, pp. 30-31).

Embedded AI does not have any visual representation. It is found in applications such as search engines or GPS. Many laboratory studies have shown that high initial trust tends to decrease as a result of AI malfunction, and that the trust recovery process requires a long time (Glikson & Woolley, 2020, pp. 31-33). In matters of tangibility, the research is limited since the awareness of the use of AI in this case is not clear. Its built-in nature suggests that people are not always aware they are using an application that uses an algorithm (Glikson & Woolley, 2020, p. 33). Although the transparency of how algorithms work generally seems to increase trust, there are studies that prove the opposite (Glikson & Woolley, 2020, pp. 33-35). It is important to note that this type of AI creates controversy and questions among users, thus undermining trust. Reliability plays an important role in embedded AI, while low reliability significantly reduces trust. Regaining it is a difficult and time-consuming process (Glikson & Woolley, 2020, pp. 35-36). Concerning task characteristics that require social intelligence, trust in people is higher than in AI. Here the research findings show that the subjective value of human self-confidence plays an important role in trust, as people who consider themselves more capable than a machine rely less on technology (Glikson & Woolley, 2020, p. 36). Finally, focusing on the immediacy behaviors of embedded AI, one can either highlight its ability of constantly monitoring users that leads to a decrease of trust or point to its personalization that increases trust (Glikson & Woolley, 2020, pp. 37-38).

2.2 Literature on the Black Box

The term black box dates back to the 17th century, but its current meaning and use can be traced back to the cyberneticians of the 1950s and 1960s. In the 17th century, the black box was allegedly used to refer to coffins, denoting the mystery surrounding their inner space; the mystery surrounding death. Additionally, when it was used in 19th century deep-sea science or, later, in the case of airplanes[3], the black box (actually an orange box as far as airplanes are concerned) preserved the aura of mystery as inputs and outputs were the only visible effects of it, whereas its construction and inner workings remained mysterious (Alaniz, 2020, pp. 596-602). Cyberneticians drew upon this metaphor in order to describe complex systems, the inner workings of which were largely under a veil of mystery (Petrick, 2020, pp. 575-582). Their work laid the foundations for the current use of this metaphor in STS and, especially, in critical studies of AI.

[3] The "black box" of airplanes was characterized as such because it initially worked like a camera obscura, i.e., it should be dark within in order to store the necessary information. Engber, D. (2014). Who Made That Black Box? The New York Times.

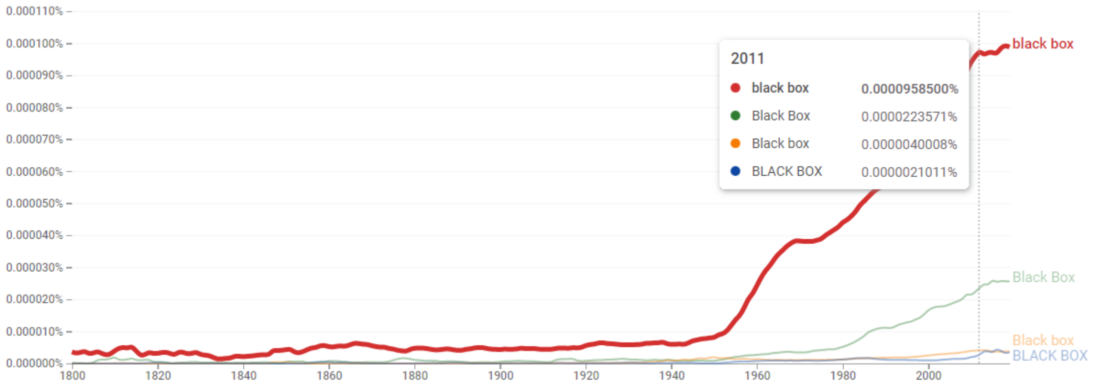

Figure 3: Google Ngram for the term “black box”.

The above Ngram shows the use of the word “black box” evolved based on data collected from Google Books results of the term. Around the 1940s, the use of the term began to increase at a very high rate. From the previous Ngrams, one can see that the use of the terms “deep learning” and “big data” is starting to increase at a very high pace, as the increase in the use of the term “black box" stabilizes. In other words, this constant increase in the usage of all of those terms, around 2009-2011, could reveal a connection between the black box rhetoric and AI.

Latour borrowed the term black box from the cyberneticians and broadened its use in order to construct his technoscientific critique. In *Science in Action*, he mentions that cyberneticians use the term “whenever a piece of machinery or a set of commands is too complex. In its place they draw a little box about which they need to know nothing but its input and output” (Latour, 1987, pp. 2-4). In *Pandora's Hope: Essays on the Reality of Science Studies* Latour uses the definition we presented in the introduction of the paper, “…the more science and technology succeed, the more opaque and obscure they become” (Latour, 1999, p. 304). His solution to the black box problem is to trace the work of the scientists in the laboratory (Latour & Woolgar, 1986, pp. 242-244). Backpropagating to the initial circumstances in which the black box was created, one may be able to open it and understand the black boxing procedures taking place. However, this is not the only way to open the black box. Marcheselli in her article “The Shadow Biosphere Hypothesis: Non-knowledge in Emerging Disciplines” challenges Latour’s schema, looking only to the present and past of scientific practice, attempting to add a third dimension, the future. Imagining a future in which the black box is fully opened could be another approach to opening Pandora’s box. Her case in point comes from challenging the very definition of life in an astrobiological setting (Marcheselli, 2020, The Shadow Biosphere: Thinking Outside the (Black) Box section). Additionally, Caitlin Wylie's paper "Glass-boxing Science: Laboratory Work on Display in Museums" shows how the black box can actually be hidden in plain sight (Wylie, 2019). Using a glass wall to make laboratories appear transparent to the public does not open the black box of science in any essential way. The scientific artifacts and procedures remain elusive (Wylie, 2019, Seeing without Understanding section).

By searching on JSTOR's database, one can see the proliferation of works around the term black box. From around 600 articles in 2014 in the discipline "History of Science & Technology" (Shindell, 2020, pp. 567-572), in 2021 it yields 4,818 results and 2,576 in "Science & Technology Studies". In this brief literature review, a selection of very few writings on the black box was made in order to outline its historical development and consider the use of the metaphor in the case of AI.

Pasquale in his book *The Black Box Society: the Secret Algorithms that Control Money and Information* elaborated on this metaphor by creating a holistic view of society that becomes black boxed based on the mystery revolving around algorithms and AI systems integrated in it (Pasquale, 2015). Through a complex network of relations and a co-construction process, the black box of AI is constituted in a reciprocal relationship with society's black box. In an article about deep-sea science and its black box, different kinds of black boxes are described in order to analyze how scientists deal with uncertainty (Pasquale, 2015). In a similar vein, AI's black box has a great deal to do with uncertainty and how to interpret it. Even if one supposes that uncertainty cannot be completely eliminated, the elucidation of the black box that is AI is important in dealing with it. In principle, an attempt to open the black box should be made in order to foster transparency, facilitate scientific practice and ensure scientific integrity.

However, dealing with uncertainty is by itself an important part of modern-day science. In the article "What good is a black box?", the uncertainty of the black box can be advantageous especially when it includes "low-risk, high-reward situations, when the cost of a wrong answer is trivial relative to the value of a right answer or when the black box objectively outperforms all other means of data analysis, even the judgment of a trained human" (Fahrenkamp-Uppenbrink, 2019, pp. 16-40). Moreover, black box "algorithms can also reveal new connections within datasets that are not intuitive from first principles” (Fahrenkamp-Uppenbrink, 2019, pp. 16-40). It concludes that although “they must be used with care, black boxes have a clear role in advancing science and engineering” (Fahrenkamp-Uppenbrink, 2019, pp. 16-40). Essentially, it is stated that the benefits of using the black box in some situations, without even trying to pry it open, outweigh the costs of the uncertainty embedded in it. However, these only include low-risk situations. But AI applications can be high-risk, as in the case of self-driving cars. It is not intuitive how the black box’s usefulness can outweigh the risks posed by its opaqueness in this case, to say the least.

Finally, there is some critique upon the very black box of STS. One manifestation of it comes from the article “White Space and Dark Matter: Prying Open the Black Box of STS”, where it is stated that white space, white-dominated technoscience, forms the foundations upon which the black box of STS is built (Mascarenhas, 2018, The White Space section). Black people are usually excluded from these studies as they constitute marginalized groups in the technoscientific paradigm. Undoubtedly, the application of the black box metaphor within STS should also be a point of self-critique for this diverse discipline.

2.3 STS Literature and Issues Regarding the AI-Black Box Relationship

STS literature has actually considered in detail the broad range of the black box concept in AI and has critically reflected on how the increase in opaqueness is positively correlated with an increase in technoscientific complexity. Latour’s definition of the black box points exactly to this attribute (Latour, 1999, p. 304). For our purposes, deep learning constitutes the primary case as it represents a cutting-edge technology in AI, where the complexity of the system reaches its peak. In the following short review, we will present some of the basic tenants of STS literature as far as the relationship between AI and blackboxing is concerned, along with the connection between black box and white box, as it can be very helpful in unpacking AI’s black box.

2.3.1 Big Data Black Box

Let us assume that the biases that occur during the collection of data are mostly related to unstructured data. When it comes to structured data sets, these may be masked, making it very difficult to actually understand their existence and impact. One could see this process as a kind of blackboxing. To elaborate a bit, through data extraction, unstructured data is transformed into a structured one. Schematically presented, this occurs through segmentation (i.e., finding “the starting and ending boundaries of the text snippets that will fill a database field) classification, i.e., determining “which database field is the correct destination for each text segment”, association which “determines which fields belong together in the same record”, normalization, i.e., putting “information in a standard format in which it can be reliably compared” and deduplication which “collapses redundant information so you don’t get duplicate records in your database” (McCallum, 2005, pp. 48-55). Thus, the structured data set appears neat and tidy. Various gender, race and other biases that were probably embedded in unstructured data seem to be eliminated. But is this so? Apparently not. Structuring data does not actually eliminate biases. It masks them through a process of blackboxing. For instance, data collected through a racial prism now appears to be normalized so as to be reliably compared with other data.

Data coming from social media users can be utilized to create personalized marketing content, to segment the market according to certain characteristics, to optimize research results and so on. Knowing that the above biases are embedded in the very process of creating this type of data and that they actually affect usage is important in order to unpack the black box of creating data in this case. But how is it a black box? Having a very high number of users, social media are perceived as instruments of democratization as they can give voice, accessibly and cheaply, to every person in the world. However, political, economic and social factors actually hinder such an aspiration. Many countries and, of course, a large number of people do not have access to the technological means necessary to use social media. Additionally, some countries have policies that ban its citizens from social media. Even among social media users, many are banned due to censorship or due to not conforming with the rules of conduct. From those left, only a few have a disproportionate contribution to the creation of content. In this sense, the blackboxing of media content creation emerges by omission—i.e., despite social media aspirations for openness and inclusion, many people are left out of its operations (Pasquale, 2015, pp. 72–96).

Furthermore, we can see how the black box of big data can emerge through a racial bias against black people, which was and still is largely embedded in US society and has been reflected on police arrests (Leslie, 2020, pp. 12-23). Of course, discrimination against black people has affected their economic and social status. Thus, more black people were likely to occupy poor households. Generally, crime and, namely, petty criminal offenses are also related to economic status. Thus, police arrests in the US were disproportionately more for black people. Amazon’s Rekognition used police databases to feed its deep learning machine (Leslie, 2020, pp. 12-23). As most criminal investigators know, one of the best, if not the sole reliable factor for predicting future criminal offences, is past offenses (Leslie, 2020, pp. 12-23). Here we have a racial bias, which was embedded in the data feeding the machine and followingly biased usage of this data as its results confirmed the expectation of prejudiced, even subconsciously, policing. Social upheaval made the opening of this black box a necessity and, finally, a reality.

2.3.2 Deep Learning Black Box

After this brief overview of deep learning, we will proceed into the processes that constitute its black box. It should be apparent by now that the discussion is not about something entirely new but about elaborate algorithms with multiple layers and interconnections that create the deep learning neural networks we know today. When applying actor-network theory to deep learning applications, we should distinguish at a high level between the programmer or the developer and the end user. What is peculiar with blackboxing procedures concerning deep learning lies in the uncertainties that expand its opaqueness further, reaching even its very developer. We will illustrate this point as we proceed through this chapter. Needless to say, deep learning’s black box is notoriously difficult to crack open. An actor that plays a critical role in the unpacking process is the end user, i.e., society. However counterintuitive this might be prima facie, we will make our case by explaining the ethical but, also, sociotechnical mandate of expanding inclusivity. Citizen science is very important in this case.

Starting from the latter, opening the algorithm to more users is feasible with today’s technology and would require simple training, along with making deep learning networks more user-friendly. As mentioned by Eric Siegel, it is relatively easy to provide end users with the means to experiment with the algorithm and explore various outcomes leading probably to pinpointing any biases or mistakes related to it (Siegel, 2016). This way, one can view the constitution of the deep learning black box from its endpoint. However, there are also other ways to understand the construction of this black box; from a technical analysis of each node to the creation of other neural networks that could work as white boxes providing access to the obscure procedures of deep learning neural networks. Moreover, the black boxing procedures taking place in deep learning are not only technoscientific but, also, social and political. Understanding the decisions that lead to the selection of a certain algorithm and its goal can reveal the black box's constitution. Moreover, one could follow Latour's way of probing into the scientific laboratory by means of learning how to program or taking advantage of already existing scientific controversies.

2.3.3 Deep Learning White Box

A quite amazing use, yet reasonable from a mechanistic point of view, of artificial neural networks or ANNs is related to improving our understanding of biological neural processes. To put it into context, deep learning neural networks can be used as white boxes in order to mirror the black box of the brain, the most elusive of them all. Despite our presentation of deep learning as a black box and the explanation of the difficulties one faces when she attempts to open it, the even opaquer neural processes that take place in the brain make it appear transparent. In this section, we will briefly explore the importance of using ANNs as white boxes and its implication not only for biology but, also, for our understanding of the workings of these very networks. However, the STS literature on the matter up to now is scarce.

Using deep learning "machines" to explain organic processes is another instance of a tradition that dates back to Descartes and the mechanistic point of view. Given that the brain is a machine and the most complex one for that matter, is still pervasive in modern scientific thought. However, there are many philosophers who have challenged this idea. Approaching biological phenomena from a mechanistic point of view requires a reduction of the former to the physicochemical domain. Arguably, the mechanical domain is more accessible to science and the prospect of using it in order to enter the biological one is still appealing. But this mindset may have serious limitations.

For instance, Canguilhem points to the fact that our point of departure should be the biological, being logically and chronologically prior to the mechanical (Canguilhem, 2008, p. 85). However, the implementation of this may seem quite difficult on a practical level. One should change one's mindset in order to overcome any limitations the mechanistic view may hold. Nonetheless, there are instances where our understanding of biological procedures can in concreto illuminate mechanical procedures. In the case of deep learning, for instance, adopting a behavioristic approach (from psychology and biology) can be used to elucidate the workings of machine learning. Although this is not an example of a completely transparent white box coming from the biological domain to open the mechanical black box, it could help understand the blackboxing procedures that construct the black box of the machine.

2.4 Issues- Research Questions from 2.2

The main research questions that constitute the frame of analysis for the primary data collected are listed below. We will attempt to answer these by reviewing the relevant literature related to the connection between AI (that is, machine learning, deep learning and Big Data) as presented in the chosen scientific journals. What are the processes through which the black box of AI is constructed and is it possible to unpack it? Should we try to open the black box of AI? Building trust in AI: what are the societal and ethical implications of opening AI's black box? Black boxes and white boxes: how does the black box metaphor penetrate biological and mechanical fields?

3.Theoretical Framework

3.1 STS: A Brief Introduction

This chapter offers a brief introduction to science and technology studies. First, it is important to mention that STS is a broad field of study. At its core, one can find the theory of social construction of technology and the related concept of co-production. Another basic STS framework is actor-network theory (ANT). Additionally, the multilevel perspective is another significant part of STS's methodology. One could also argue that the citizen science perspective is a crucial part of STS and an expansion to the social constructivism approach to science. Of course, there are many other frameworks that are used in STS and, for our purposes, we will use the theoretical perspective of the black box concept in order to analyze AI technologies and their impact to society, science and technology (Felt, 2017, pp. 41-49). Before delving into the nuances of this concept, the black box metaphor, though, we will present the development of STS. Despite its relatively short history and the newly drawn attention to it, an overview of its historical development can be very instructive. By doing so, the various frameworks of STS are going to come into view.

The field of STS is a relatively new field of study that is based on interdisciplinarity. It emerged through the growing interest regarding the interplay between science, technology and society. The so-called sociology of scientific knowledge or SSK, which, was developed alongside STS was also trying to analyze the sociological point of view regarding science and technology. The sociology of scientific knowledge, whether one talks about the more theoretical *strong programme* of the school of Edinburgh or the more empirical studies of the school of Bath, is based on symmetry and the avoidance of a sociology of error. Scientific controversies and their resolutions are not seen through a prism of right-versus-wrong visions of science but as points in scientific history where scientific practice becomes transparent or enters a black box phase (Pinch, 2015, pp. 281-286). STS, with key proponents such as Latour, Woolgar, Knorr-Cetina, and Lynch, among others, calls for a more localized analysis of the interplay between science, technology and society (Mukerji, 2001, pp. 13687–13691). Moreover, it focuses on how technoscientific artifacts are co-constructed or co-produced through a sociotechnical process.

Based on the influential work of Thomas Kuhn, *The structure of scientific revolutions*, STS tried to exploit this historicist turn in the philosophy of science in order to analyze scientific practice within a certain paradigm. Focusing on a specific point of view of Kuhn's work, STS scholars, Latour among others, pointed to the scientific controversies that disrupt normal science, which helped the latter formulate his theory of the black box concept along with actor-network theory (Felt, 2017, pp. 266-272). Schematically presented, the black box of science is opaque during the normal science period, but scientific controversies can be utilized to elucidate this black box. Following scientists at the laboratory is one of Latour's propositions in order to open the black box of science by tracing its origins, the moment of construction, or using loosely Kuhn's terminology, the moment when scientific practice becomes normalized (Felt, 2017, pp. 266-272).

Starting with actor-network theory (ANT), Latour wanted to point out how scientific practice is developed within a large network of actors that are closely related to each other. Every actor within the network is shaped by her relations to other actors or points in the network. Being essentially relational, ANT draws from the post-structuralist tradition to describe the fluidity of the network. In this framework, the micro and macro levels of analysis are not distinct but emerge through a relational network of various actors. The similarities or differences of the actors do not predate the network but are products of the relations built within it. ANT is a very useful conceptual tool to analyze the heterogeneous relations that emerge within a scientific, sociotechnical network (Felt, 2017, pp. 41-49).

The social construction of technology (SCOT) is a key approach for STS. With terms such as co-production and co-construction, one can understand how this interdisciplinary field views technology and science within society. More specifically, the sociotechnical co-production reveals the reciprocal relationship between technoscience and society. It shows how technology shapes and is shaped by various social groups, along with the variety in significance assigned to it by different groups. By analyzing the relevant groups and the construction of technoscience in a social setting, SCOT provides us with an interpretation of how the black box of technoscience is constructed (Bijker, 2015, pp. 135-140). However, this is also a point of criticism. Bringing to light the construction of this black box may leave aside its implications for society after this construction. It may also avoid making any axiological statement, fail to account for alternatives that did not make it or present the whole picture of the sociocultural factors that come into play in this co-production (Felt, 2017, pp. 41-49). Nonetheless, in response to such criticism, STS scholars point to the fact that adopting a social-constructivist approach actually reveals the ethical and social implications of this blackboxing process. It would help to understand the ethics embedded in the technological design and attempt to increase transparency by opening the black box in its initial stages.

Additionally, STS adopts a multi-level perspective, which refers to analyzing a sociotechnical artifact in many different levels, from the scientific practice to its cultural significance, relevant social groups and so on. One can understand this approach by implementing the relational ontology framework when talking about an ontological multiplicity. In other words, a technoscientific artifact can have many ontologies, many lives one could say, arising from its relations with other artifacts, social groups or actors within the broader sociotechnical network. The multi-level perspective can also be implemented to gain the broader picture by shifting perspectives from bottom-up to top-down. For instance, on the one hand, one can use it to analyze grassroots innovations and talk about policy and top-down governance of innovation on the other (Felt, 2017, pp. 41-49).

Another aspect of STS worth mentioning (but probably controversial within the field) is citizen science. STS scholarship raises the mandate for increasing inclusivity and transparency in the co-production of sociotechnical artifacts. Citizen science, becoming possible by today's technologies to disseminate information, can be seen as an approach within STS that aims to fulfill some of its mandates for instance, to increase inclusivity and, thus, accountability (Kimura & Kinchy, 2016, pp. 331-345). For our purposes, citizen science can help open the black box of AI by focusing on the end user, starting from the endpoint to follow it back to the construction of the black box. This is quite the opposite of what Latour wanted to do. Let us illustrate our point. Suppose one has a deep learning algorithm that can provide too many results, which make it virtually impossible to test the various combinations and account for any biases or misgivings. However, making it user-friendly and providing people with a basic education will make the use of this algorithm possible for more people, who can test various combinations and get very different results. Increasing inclusivity in this way may help increase accountability as more people will have access to the algorithm and will probably be able to pinpoint any biases or flaws in it.

To conclude, STS is a diverse, interdisciplinary field with various methodological approaches at its disposal. In this brief presentation, one could mention only a few, albeit the most prominent ones and those relevant to the purposes of the thesis (Felt, 2017, pp. 41-49). In the next section, the black box metaphor, the main STS methodological tool used in this thesis, will be presented.

3.2 Black Box

For the purposes of this thesis, the black box metaphor is considered to be the appropriate methodological tool to analyze AI and its black box from an STS perspective. Undoubtedly, this metaphor cuts through several STS theoretical frameworks, such as the multi-level perspective, ANT, and SCOT to mention a few (Felt, 2017, pp. 41-49). Having already talked about the origins of the metaphor, within STS, Latour was the first to turn the black box metaphor into a methodological approach that can be used to interpret scientific practices. Understanding the black box of science, the necessity to pry it open and the ways in which it is possible to unpack it are very important for the STS critique. All the more so when it comes to AI and its pervasive black box.

The main issues regarding the black box metaphor and its use in STS have already been discussed. Next, we turn to the application of this metaphor as a methodological tool to describe and analyze technoscientific and societal problems. Again, understanding the co-construction of AI's black box, the necessity to open it and ways in which it can be opened could be very enlightening. Creating or using transparent white boxes to open black ones creates a framework of understanding, which, apart from making a description of the fact, provides critical tools to improve the situation. The critique to the co-construction of the black box and to the very concept of it dates back to 1993 in Winner's article "Upon Opening the Black Box and Finding It Empty: Social Constructivism and the Philosophy of Technology", where he argues that the black box of social constructivism is filled with vague, abstract notions (Winner, 1993, pp. 362-375). Instead, he proposed a more technical, specialized approach to the problem (Winner, 1993, pp. 362-375). However, the responses of STS scholars were many and, generally speaking, the view of extreme specialization in a field is somewhat contrary to the modern-day mandate for interdisciplinarity. The purpose is to facilitate discourse between experts and not to isolate them. One of those responses, directly pointing to this article but not exactly a one-on-one response, comes from Steen's article "Upon Opening the Black Box and Finding It Full: Exploring the Ethics in Design Practices", where it is argued that the ethics, along with the sociopolitical issues involved, are of paramount importance for the analysis of technoscientific discourse (Steen, 2015, pp. 389-412).

4. Methodology

4.1 Secondary Literature

The secondary literature includes various historical and philosophical writings on AI. The purpose was to present an overview of the framework through which one should view AI and the modern debate around it. More specifically, when it came to literature regarding the black box and the relationship between the black box metaphor and AI, the research focused on STS journals, books and articles. From Latour's and Pasquale's books on the black box concept to STS journals such as *Social Studies of Science, Science & Technology Studies: Journal of the European Association for the Study of Science and Technology,* and, mostly, *Science, Technology & Human Values* in order to find the relevant articles for this thesis. The keywords for our research included the terms: “black box”,” Big Data black box”, “Deep learning black box”, “AI black box”, and “machine learning black box”.

4.2 Primary Literature

Researching the scientific journals *Scientific American* and *Nature* for articles relevant to our research questions, we used the following keywords: “AI black box”, “machine learning black box”, “deep learning black box”, “big data black box” and “black box”. We also tried different combinations of the terms “black box”, “deep learning”, “machine learning” and “big data”, but the above keywords were the ones yielding optimal results. Apparently, the research engine in these journals was somewhat case sensitive. However, in most cases, the results did not change when the words were capitalized, with the exception of big data.

Going through the results, we used some criteria to help us collect the necessary articles for this thesis. Firstly, sometimes the research engine brought back many pages of results, but one could see that the relevance of the titles to our research questions was almost zero after a certain point. That point was often reached somewhere after page ten of the results. Thus, in none of these searches did we reach beyond the twentieth page. Within these results, very few of the articles’ titles contained our basic keywords, whereas in some cases these keywords, such as “black box”, had a completely different meaning. More specifically, we searched for the term “black box” as a metaphor, a methodological tool to understand AI and its applications. However, in some articles, the term “black box” referred to the black box of an airplane. Despite the fact that there is some loose connection to the airplane’s black box and our use of the term, these articles were completely irrelevant yet easy to rule out from their very title.

Secondly, the articles collected were published within the last decade. We aimed to avoid articles that were too old to offer insights into modern views of AI but, at the same time, to gain an overview of AI’s recent boom and its critique. Of course, the discussion around AI’s black box is not very new, but its intensity has risen exponentially in the last decade, and it seems to continue at the same pace. Moreover, some of the articles do not explicitly mention the terms in question but they were chosen because of their thematic relevance to our research questions. Arguably, the articles about trust in AI or, more generally, those within the field of ethics, even if they seem to go beyond the STS domain, strictly defined, addressed issues, such as transparency, directly related to the black box of AI and the mandate to open it. For these reasons, diverse articles were selected and interpreted with the appropriate caution, always within the framework of the black box metaphor.

Given the use of the aforementioned keywords and their combinations, the exact replication of the study may be hindered by the implication of our relevance criterion. This is why it is imperative that we analyze how the collection by relevance took place. First, the titles of the articles were considered in order to find words related to the black box concept, such as opaqueness, interpretability and the like. Moreover, ethical problems and issues regarding transparency are closely linked to AI's black box. More general concepts related to AI, machine learning, deep learning and Big Data were also researched in the titles and the content of the articles via the prism of providing us with a discussion about the historical development of AI and the basic questions relevant to its capabilities and problems. The reason for this has to do with creating a better understanding of how AI's promises, whether it actually stood up to them or not, improve our realistic insights about AI and the procedures related to it. The purpose was to find papers that critically reflect on our current understanding of AI and its potential. Thus, many articles revolve around both current and future applications of AI.

Finally, we chose to limit our research to *Scientific American* and *Nature* because, on the one hand, they provided us with the necessary popular science articles for our thesis and, on the other, expanding our study to more journals would make it more difficult to process. Additionally, access issues contributed to the selection of the above journals. Time restrictions also played a role in the decision to involve only two scientific journals in our study. However, we believe that the main themes related to the discussion around AI and its black box are illustrated in detail in the articles collected from these journals. The inclusion of more articles would probably not have a core but a peripheral effect on the expansion of the thesis.

5. Primary Data Presentation

In this section, we present the primary data collected. Through the discussion of articles from *Nature* and *Scientific American* we create a comprehensive overview of how the content that is relevant to our research questions is depicted in these journals. We begin with *Nature's* articles followed by those of *Scientific American.* Starting by Castelvecchi's article, which appears in both, the foundations are built in order to process the rest of the articles about the relationship between the black box and AI and some issues related to various perspectives on AI. Ethical issues regarding AI are also presented as they are of great concern nowadays and they interact with the black box concept, at the crossroads of transparency for instance. Trust in AI is closely related to transparency and, of course, the black box problem.

In his article, back in 2016, Davide Castelvecchi straightforwardly addresses the big question: "Can we open the black box of AI?" (Castelvecchi, 2016). Undoubtedly, this article is very important for the purposes of this thesis. Starting from the conclusion, we will then endeavor to tackle the issues presented in the main body of the article, which are, arguably, much more poignant. Castelvecchi concludes with Pierre Baldi's comment, currently a distinguished professor at the Department of Computer Science of UCI, that scientists should proceed with their work without being "too anal" about the black box (Castelvecchi, 2016, p. 23). They actually carry one in their heads all the time, the hardest to crack of them all. Of course, this comment about the brain is Baldi's final assertion towards the practical significance of opening the black box. It seems that on balance, trying to pry open the black box of AI can be cumbersome and even hinder scientific practice. It goes without saying that scientific practice should proceed without any hindrances. Nonetheless, scientists can never be too analytical about AI's black box as it presents a very good opportunity to critically reflect on their practice, analyze its consequences and understand its limitations. The analogy between the brain and AI machines is very interesting and actually encompasses the whole article. But now it is more relevant to present an overview of the main points from Castelvecchi's article related to AI's black box.

The article starts with Dean Pomerleau's initial encounter with AI's black box. A robotics graduate student at Carnegie Mellon University in Pittsburgh, Pennsylvania, back then, Pomerleau trained a Humvee military vehicle to drive itself. When the vehicle encountered a bridge, Pomerleau had to quickly grab the wheel in order to avoid the crash. The obvious question he had to ask is "what went wrong?" In doing so, he understood that the neural network he created constituted a black box. At the time of Castelvecchi's article, twenty-five years have passed since Pomerleau's first attempt to understand and analyze the black box of AI within the confines of a system with limited capacity. Modern advancements in technology, such as deep learning neural networks, have largely exacerbated the situation, making the black box problem “all the more acute” (Castelvecchi, 2016, p. 22). Without a doubt, this situation still stands and even more so today, five years after Castelvecchi’s article.

Some of the questions raised by the situation above include: “how exactly does an AI application, for instance a deep learning one, reach a specific result?”, “how can anyone be sure the results are correct?” and “how far should one go to trust AI?” (Castelvecchi, 2016, pp. 21-23). Scientists nowadays proceed just as Pomerleau did in the case of the self-driven vehicle; they try to open the black box of AI. Just like a neuroscientist, they try to analyze neural networks in order to find the specific connections that lead to the resulting answer given by the machine. Distinguishing between different answers is not enough for a scientist who wants to know the exact characteristics of this difference. As indicated in the deep learning section above, the first neural networks were created in order to simulate the “neurons” of the brain (Castelvecchi, 2016, pp. 21-23). These “digital synapses” start from the bottom layer, where simple correlations between the synapse and the object are manifested, and they reach the top layers through multiple processes, where certain results are reached, such as the classification of objects (Castelvecchi, 2016, pp. 21-23). In order to penetrate these processes and open the black box of deep learning in the process, scientists have used techniques like “Deep Dream”, where a specific characteristic is manifested at a disproportionally higher frequency than others (Castelvecchi, 2016, p. 22). This way, they can understand how important the specific characteristic is for the neural network.

A brief passage from Castelvecchi’s article summarizes how machine learning and, especially deep learning instances, work. As the author says:

> The power of such networks stems from their ability to learn. Given a training set of data accompanied by the right answers, they can progressively improve their performance by tweaking the strength of each connection until their top-level outputs are also correct. This process, which simulates how the brain learns by strengthening or weakening synapses, eventually produces a network that can successfully classify new data that were not part of its training set. (Castelvecchi, 2016, p. 22)

Again, the pervasive analogy from the biological to the mechanical is manifested.

Deep learning networks need big data —i.e., huge amounts of information that they can process in order to reach a specific result. Given that humans are not capable of processing this volume of information, the black box of AI thickens. One example mentioned in the article has to do with the hypothetical scenario in which a machine trained by a large number of old mammograms identifies an apparently healthy woman as cancer-stricken (Castelvecchi, 2016, pp. 22-23). The work of the physician actually becomes more difficult as they cannot pinpoint the reasons for this result. If the machine cannot explain itself and the human interpreter has no access to its black box, then scientific practice is hindered. As Michael Tyka, a biophysicist and programmer at Google in Seattle, Washington, put it: “The problem is that the knowledge gets baked into the network, rather than into us…” (Castelvecchi, 2016, p. 23). He goes on to say: “Have we really understood anything? Not really—the network has” (Castelvecchi, 2016, p. 23). Obviously, glimpsing at how the black box is constructed is not enough and the mandate to open it becomes more urgent than ever, especially in the case of biomedicine.

One attempt at opening the black box of AI is reverse-engineering the machine’s workings. Andrea Vedaldi, a computer scientist at the University of Oxford, UK, and his group tried to do so by taking the algorithms that Geoffrey Hinton, a machine-learning specialist at the University of Toronto in Canada, had developed in order to improve neural network training and running them in reverse (Castelvecchi, 2016, pp. 22-23). They tried to reconstruct the images that were represented by the neural network. Following this process, one can reach an understanding of the importance assigned by the machine to various features. Deep Dream works in the same way (Castelvecchi, 2016, pp. 21-22). Moreover, its online distribution gives access to many end users, who are given the opportunity to explore ways in which an AI application can be reverse-engineered.

Being notoriously difficult, opening deep learning’s black box has led some scientists and scholars to believe that deep learning is not the answer to many world problems and that simpler solutions should be adopted in order to increase transparency in the scientific process (Castelvecchi, 2016, pp. 22-23). Although increasing transparency comes hand in hand with opening AI’s black box, the mandate for the former is not to be at the expense of the latter. Simpler, more transparent models should not act as substitutes for complex deep learning applications, but they should supplement them as the world is full of complex problems that may not be susceptible to reductive solutions. In other words, increasing transparency is a valid aspiration and a motive to open the black box of AI, but reducing complexity in order to do so may simply transpose the problem rather than actually solve it. This is why, as articulated by Castelvecchi, reductive models should be complementary to complex deep learning ones (Castelvecchi, 2016, pp. 22-23). However, transparency remains critical for scientific practice and this is why the black box problem should be continuously addressed.

The article by Castelvecchi, as presented above, forms the basis for our discussion of *Nature* and *Scientific American* articles, as it appears in both when one searches for the relationship between AI and the "black box" concept. More articles from *Nature* address issues related to the black box theme and trust-in-AI issues related to it. For example, in "Learning from the machine", Haiman points to the fact that the deep learning black box in cosmological applications may not be a concern for big companies, but it is an issue when it comes to trust-in-AI results (Haiman, 2019, pp. 18-19). Statistics from cosmological datasets using deep learning techniques have proven to be more effective than human statistical analyses. Yet, being able to "understand the physical origin of the information" is a crucial element in building trust in AI systems, and this is why the black box problem should be also addressed thoroughly in this case (Haiman, 2019, pp. 18-19). In "Illuminating the dark side of machine learning", it is stated that "one major challenge is that the machine-learned relationships often remain enigmatic in the model; this 'black-box' nature hinders the elucidation of the actual biological mechanisms determining the output phenotypes", which raises the black box problem and the relevant transparency issues in the field of genome sequencing and its biomedical applications (Burgess, 2019, pp. 374-375).

Addressing the issue of trust in AI can have a very wide range, but in the next article one aspect of it comes into focus. Specifically, in the manufacturing industry, where AI can use big data in a factory to improve the efficiency of the production process, reduce the energy consumption or even predict failure of devices, trust in AI can be seen as an important factor that impacts implementation. As mentioned in the article, a typical AI-based predictive maintenance can reduce annual maintenance costs by 10%, unplanned downtime by 25% and inspection costs by 25% (Li et al., 2021, p. 13564). Despite the benefits that can be provided by AI in the specific industry, a recent survey shows that 42% of people lack basic trust in AI and 49% cannot even name an AI product they can trust (Li et al., 2021, p. 13564).

A black box can be, as already mentioned, a deep learning system that is comprised of complex neural networks modelled on the human brain. The result is that the black box is impenetrable to humans, while there is a great need to explain the relevant AI processes, especially in fields like healthcare or transportation, where lives may be at stake. In contrast to this situation, as explained in the article, wide learning works on a simple principle: that is, “evaluate every combination of data item in a system to find important combinations, called knowledge chunks” (“Building natural trust in artificial intelligence”, 2021). Using this approach “can generate accurate predictions even with a small volume of training data” (“Building natural trust in artificial intelligence”, 2021).

Additionally, a problem with AI systems is that their accuracy can deteriorate significantly over time. For instance, a financial system built on training data that has become outdated due to new market conditions will not produce accurate results (“Building natural trust in artificial intelligence”, 2021). This is where high durability learning comes into play. It is the world’s first technology that “can automatically estimate the accuracy of an AI system and, if necessary, update so its predictions remain valid” (“Building natural trust in artificial intelligence”, 2021). This type of learning works by comparing “the distribution of data from when an AI system was trained and the distribution changes of data from actual operations” (“Building natural trust in artificial intelligence”, 2021). By analyzing differences in those datasets, it can quantitatively estimate the system’s accuracy. Also, high durability learning can “automatically adapt an AI system to new input data, maintaining accuracy without expensive retraining” (“Building natural trust in artificial intelligence”, 2021). The effects of high durability training on the black box of AI have to do with the processes involved in maintaining system accuracy. Given that accuracy is measured by the results the AI system provides, this type of training is bound to interfere with interior processes in ways that are not known in advance at their entirety.

Ethical concerns regarding AI are a huge issue that is presented in many articles nowadays. It is closely linked to the black box concept as AI systems may be used to “amplify discrimination and biases, such as gender or racial discrimination, because those are present in the data the technology is trained on, reflecting people's behaviour" (Castelvecchi, 2019). This passage reflects big data's black box, which presents a problem that raises ethical concerns. Ultimately, AI ethics are a projection of human ethics to the machine. As explained by Lo Piano, "the set of decision rules underpinning the AI algorithm derives from human-made assumptions, such as, where to define the boundary between action and no action, between different possible choices" (Lo Piano, 2020, p. 5). Of course, "this can only take place at the human/non-human interface: the response of the algorithm is driven by these human-made assumptions and selection rules" (Lo Piano, 2020, p. 5). Concerning big data's black box, it is argued that even "the data on which an algorithm is trained on are not an objective truth, they are dependent upon the context in which they have been produced" (Lo Piano, 2020, p. 5). In the same article, the construction of machine learning instances as black boxes is also addressed and issues regarding transparency, accountability and fairness are raised and analyzed critically via case studies taken from the application of machine learning in the judicial system and autonomous cars (Lo Piano, 2020, pp. 1-6). The constant and rapid evolution of machine learning models, which learn and improve constantly, presents urgent ethical issues as the more they learn, the harder it is to understand them; the opaquer their black box gets (Lo Piano, 2020, pp. 5-6).

After establishing the link between transparency and the black box framework through a brief description of the above articles, one can now attend to another article that presents the case of a deep learning algorithm for glaucoma. As stated in the article, this algorithm is a black box, and in order to understand how it reaches diagnosis one has to peer into it (Xu et al., 2021, pp. 1-10). One way to do so, as it is argued, is to implement a hierarchical approach to deep learning in order to increase transparency and interpretability. The researchers built a hierarchical deep learning model for a small portion of the sample, which models the thinking of experts as much as possible (Xu et al., 2021, pp. 1-10). This model could separate the variables that had a causal effect to the resulting diagnosis, thus improving the efficiency of the system by ameliorating human–machine interaction (Xu et al., 2021, pp. 1-10). In this case, one can see that expert human thinking was used as a white box in order to mirror and, eventually, pry open the black box of deep learning.

Another article pointing to AI used in interpreting medical images states that, on one hand, explainability would be desirable as it would increase trust in AI, but, on the other, complex deep learning algorithmic explanations may not be readily accepted by expert physicians (“All eyes are on AI”, 2018). However, these algorithms work within the confinements of appropriate medical practice, and they do nothing but help physicians in their work. After all, as stated, providing better care is “an ill-defined, subjective task” (“All eyes are on AI”, 2018). However, such a conclusion that would lead to the acceptance of the black box instead of trying to open it relies on a rather dubious presupposition that these algorithms can “only” help the physician and not hinder her work. But, being ignorant of their explanations, physicians may be led to contradictory or even uncalled-for diagnoses. Thus, the black box bounces back as a problem that needs to be addressed

At the intersection of the biological and the mechanical comes the black box of reprogramming stem cells. This notoriously impenetrable black box is derived from the fact that scientists have an accurate understanding of the input (differentiated cells) and the output (pluripotent cells), but they are completely ignorant of the mechanics involved in the process of reprogramming (Cyranoski, 2014, pp. 162-164). As mentioned in the article, they think that reprogramming involves deterministic (that is, a specific event has to be followed by another specific one) and stochastic (an event described by probabilistic distribution) processes (Cyranoski, 2014, pp. 162-164). However, there are scientists that think such a distinction has no value as probability can be assigned to all steps of the process. Moreover, using a stochastic model has proven that the randomness of the reprogramming process “can be controlled or even eliminated” (Cyranoski, 2014, pp. 162-164).

Cynthia Rudin’s article is opposing the mandate to explain the black box, yet addressing the same issues. According to her, a black box machine learning model could be either a function too complicated for any human to comprehend or a function that is proprietary (Rudin, 2019, pp. 206-214). In both cases, an explanation is required, which would consist of a separate model trying to replicate most of the behaviour of a black box. The writer notes that the term ‘explanation’ here refers to an understanding of how a model works, as opposed to an explanation of how the world works (Rudin, 2019, pp. 206-214). The latter adds to her concern that the field of interpretability, explainability, comprehensibility and transparency in machine learning has strayed away from the needs of real problems. This field dates at least back to the early 1990s, and there are a huge number of papers on interpretable machine learning in various fields, though they often do not have the word ‘interpretable’ or ‘explainable’ in the title, as recent papers do (Rudin, 2019, pp. 206-214). Recent work on the explainability of black boxes—rather than the interpretability of models—contains and perpetuates critical misconceptions that have generally gone unnoticed, but that can have a lasting negative impact on the widespread use of machine learning models in society (Rudin, 2019, pp. 206-214). It is very important to have in mind that a machine learning method cannot provide a completely faithful explanation of the process that the original model computes; otherwise there would be no need of the original model in the first place. This fact can be dangerous since the explanation model can lead to an inaccurate and possibly misleading representation of the original model (Rudin, 2019, pp. 206-214).

Another fact mentioned in the article is that a black box can have the characteristic of uncovering “hidden patterns” (Rudin, 2019, pp. 206-214). The data that a black box is gathering and the way they are processed can uncover hidden patterns that the user was not previously aware of. The writer argues that the creation of an interpretable model of that black box can locate these patterns and use them (Rudin, 2019, pp. 206-214). However, this interpretable model must be transparent and flexible enough to fit the data accurately, which is another difficulty that the construction of interpretable models is faced with. Despite the aforementioned issues that the interpretable models are facing, Rudin strongly encourages efforts that should be made so as to be able to create interpretable and explainable machine learning processes in order to restrict the use of black boxes in various situations where their use could lead to safety issues or poor decision making (Rudin, 2019, pp. 206-214).

These issues are also raised by Yoshua Bengio, a Turing Award winner for his work on deep learning, who has significantly contributed to the establishment of international guidelines for the ethical use of AI (Castelvecchi, 2019). Those were established, initially through the Montreal declaration in December 2018, where the ethical principles of the use of AI were increased from seven to ten, and, secondarily, through the efforts of creating an organization in Montreal, the International Observatory on the Societal Impacts of Artificial Intelligence and Digital Technologies (Castelvecchi, 2019). This organization aims at bringing together all the relevant actors: governments, because they are the ones who are going to take action; civil-society experts, which means both experts in AI technology and in the social sciences, health care and political science; and companies that are building these products. “Selfregulation is not going to work" he argues, adding that organizations should pledge to follow specific guidelines, set carefully, since some might push things in a direction that favours their bottom line (Castelvecchi, 2019).

Concerning once again the black box of big data, the article "Big Data: Stealth control" draws on Pasquale's "black box society" in order to present the case of the algorithmic black boxing of data in modern society (Aftergood, 2015, pp. 435-436). Nowadays, the proliferation of digital devices has increased inclusivity, but, at the same time, it has led to an increase in data that has to be analysed via machine learning instances. Human processing of such vast amounts of data would be too slow or even impossible in many situations. However, the black box of algorithms creates this black box society that Pasquale is talking about by raising the issues of data privacy, governance and many more societal and ethical ones. It creates a "society in which basic functions are performed in deliberate obscurity through the collection and algorithmic manipulation of personal data" (Aftergood, 2015, pp. 435-436). Overdependence on search engines, the use of machine learning software for marketing purposes, credit score algorithms and, in general, AI applications assign value to data in ways that remain obscure for the majority of the users, yet profoundly affecting everyday life in society. Pasquale is worried that "they can be used to shape what we know, how we are perceived and what opportunities we will be afforded" (Aftergood, 2015, pp. 435-436). Pasquale is optimistic that even the opaquer institutions can be held accountable. Thus, moving from the black box society to its opposite, the intelligible society, is possible (Aftergood, 2015, pp. 435-436). However, as indicated by the article, "elucidating the problem is a first step" (Aftergood, 2015, pp. 435-436).

Another article from *Nature* points to the fact that one should turn to explanatory models in order to open the black box of deep learning. Deep neural networks have become more and more useful as well as successful in a wide range of areas in real world applications. In the article, there is an issue raised, which concerns the fact that various studies have shown that these learning machines can also result to Clever-Hans-like moments—that is, human-undesired strategies where the machine exploits artifacts in the dataset (Schramowski, 2020, pp. 476-485). The writers introduce XIL (explanatory interactive learning), which is a model that adds into the training process a scientist who interacts with the machine learning process by providing feedback on its explanations. XIL relies on two assumptions: firstly, that faithful explanations can be computed and, secondly, that the user feedback is faithful too (Schramowski, 2020, pp. 476-485). Both of those assumptions are open-ended issues and are subject to further research, though the experimental use of XIL on phenotyping, as presented in the article, demonstrates a set of positive results as well as limitations (Schramowski, 2020, pp. 476-485). Overall, XIL can help us avoid Clever-Hans-like moments in machine learning and promotes the scientific discoveries that can be produced through the interaction of humans and machines, though it is still an open-ended field with many variables that require further research (Schramowski, 2020, pp. 476-485).

Regarding the above Clever-Hans analogy taken from the animal kingdom and its nuances one can find another article in which there is reference to a trial-and-error algorithm that can help damaged robots find the necessary behavior to compensate for the damage (Possati, 2020, pp. 2-3). In other words, they can adapt to change, much like animals. However, as demonstrated by the notorious Clever Hans case, without being cognizant of the cues taken in order to adapt to new behaviors, this adaptation remains largely a black box. The main purpose of the article, though, is to integrate Lacan's thinking into a rereading of Latour in order to speak about the AI unconscious. The study of "AI systems that process big data only from a mathematical and statistical point of view significantly undermines our understanding of the complexity of their functioning, hindering us from grasping the real issues that they imply" (Possati, 2020, p. 2). AI systems are constructed as black boxes that "can produce injustices, inequalities, and misunderstandings, feed prejudices and forms of discrimination, aggravate critical situations, or even create new ones" (Possati, 2020, p. 2). Two explanations are given here for the black box of AI. On one hand, "for legal and political reasons, their functioning is often not made accessible by the companies that create and use them" (Possati, 2020, p. 2). On the other, "the computation speed makes it impossible to understand not only the overall dynamics of the calculation but also the decisions that the systems make" (Possati, 2020, 2). Thus, "engineers struggle to explain why a certain algorithm has taken that action or how it will behave in another situation" (Possati, 2020, pp. 2-3).

As implied in Possati's article, Woolgar and Latour claim that to "open" a fact means to "continue discussing about it", whereas to "close" a fact means "to stop discussing" about it (Possati, 2020, p. 11). Controversy is essential for these thinkers and especially for Latour. Topics that are important and attract the attention of the researchers are continuously discussed, whereas others retreat into obscurity. The latter solidify into black boxes, but the former are reopened and analyzed. Among the facts, there is a relationship of "gravitational attraction"; a re-opened fact "attracts" other facts and forces the researchers either to reopen or to close them (Possati, 2020, pp. 11-12). Related to the psychoanalytical black box, Lacan distinguishes between two of them in the mirror stage. The first black box "coincides with the imago itself, which hides the mirror and the rest of the surrounding world" (Possati, 2020, pp. 5-7). This "imago produces the auto-recognition and the identification that are abstractions from the technical and material conditions that constitute them" (Possati, 2020, pp. 5-7). The image constituting "the child's identification is also what blinds the child" and makes her "incapable of grasping the imaginary nature" of her identification (Possati, 2020, pp. 5-7). This first black box is weak "because it closes and reopens many times" (Possati, 2020, pp. 5-7). However, the second black box is much more stable and "coincides with the transition from the imaginary to the symbolic, therefore with the Oedipus complex" (Possati, 2020, pp. 5-7). The symbolic "closes" the mirror stage, turning it into a black box. According to Lacan, the symbolic removes the imaginary, making it a "symbolized imaginary" (Possati, 2020, pp. 5-7). Reducing it to a black box, "the imaginary can be limited, removed" (Possati, 2020, pp. 5-7). This procedure "is the origin both of the distinction between conscious and unconscious, and of a new form of unconscious" (Possati, 2020, pp. 5-7). The article goes on to uncover how this unconscious is hybridized in the digital era. AI's black box is supposed to expand the human unconscious, creating a new, hybrid form of it.

Turning now to the *Scientific American* articles in order to answer our research questions, we begin with an article addressing the black box problem directly. In "Demystifying the Black Box That Is AI", issues regarding trust in AI are linked with the blackboxing processes related to it (Bleicher, 2017). As indicated, the pinnacle of AI applications, deep learning, "allows neural nets to create AI models that are too complicated or too tedious to code by hand" (Bleicher, 2017, Fine-Tuning section). These systems may be "mind-bogglingly complex, with the largest nearing one trillion parameters (knobs)" (Bleicher, 2017, Fine-Tuning section). Feeding huge amounts of data into a deep learning machine produces results that if wrong, can be tweaked towards the right ones through a feedback process. What is interesting in these lies in their ability to learn and evolve without constant human involvement. However, the problem has to do with their arising complexity, which increases opaqueness. The vast number of parameters involved in the deep learning process (i.e., vectors and synapses) makes it impossible for the human agent to gain a complete oversight of it. Being unable to understand the reasons behind the results produced by the machine decreases trust in the machine. As stated in the article, small amounts of trust in AI may not be a problem in the case of Google's AlphaGo neural net, which when "played go champion Lee Sedol last year in Seoul" "… [it] made a move that flummoxed everyone watching, even Sedol" (Bleicher, 2017, Digital Subconscious section). But in the case of a self-driving car, an unexpected "move" by the machine can be fatal. Thus, trust in AI could form the basis to address safety issues related to it. Elucidating and opening the black box of AI is crucial for building trust.

Peering under the hood of AlphaGo, for instance, and finding the exact knob could be possible in order to assign the specific numerical values to the move produced. However, this process would be redundant. In this article, it is stressed once again that the information, the "meaning" of the move, is not stored in a specific node but diffused throughout the network, much like how the brain works (Bleicher, 2017, Unmasking AI section). Evidently, researchers have to find ways to pry open the black box of AI. In Bleicher's article, two approaches are proposed. One is called the "observer" approach and the other "surgical". The former could be likened to a behaviourist approach to AI. Understanding that the AI system is a black box, one experiments with it and tries to infer its behaviour. Model induction is another name for it, as one attempts to understand the processes involved by analysing the end behaviour of the machine. The "surgical" approach "lets us actually look into the brain of the AI system," as Alan Fern puts it, a professor of electrical engineering and computer science at Oregon State University (Bleicher, 2017, Unmasking AI section). According to Fern, getting into the neural network and exerting an "honest-to-goodness explanation" would involve tracing "every single firing of every node in the network" (Bleicher, 2017, Unmasking AI section). This way, "a long, convoluted audit trail that is completely uninterpretable to a human" would be created (Bleicher, 2017, Unmasking AI section). Fern's team proposes to use another neural network to probe the target one. This explanation neural network would be a way to find meaning in the processes involved. Even if this procedure would not be as exact as analysing every single firing of a neuron in the net, it would be a much more plausible and faster way to find meaning in the processes of a deep learning neural network.

The quest for building trust in AI by opening or elucidating its black box can involve different approaches. Bonsai, "a start-up developing a new programming language called Inkling to help businesses train their own deep-learning systems to solve organizational problems" wants to change how deep neural networks learn in order to increase transparency and foster trust in the system (Bleicher, 2017, Unmasking AI section). Learning through trial and error may be the only way deep neural networks learn up to now, but Bonsai would like to model human teaching methods in order to train the machine. Joel Dudley, director of Biomedical Informatics at the Icahn School of Medicine at Mount Sinai in New York City, is not very concerned with the black box nature of Deep Patient, a deep learning neural network by his team, but he wants to demonstrate its safety "during clinical trials" (Bleicher, 2017, Unmasking AI section). Whatever the approach, the black box and trust in AI represent two closely connected problems. At the end of the article, an important issue about transparency is raised. The very nature of transparency is multifaceted as it is contingent on the agent. It may vary depending on the actor on the network, the stakeholder so to speak. For instance, transparency of an AI system is perceived very differently by a programmer in contrast to a lawyer or a layperson (Bleicher, 2017, Show and Tell section). However, as stated at the end, "being able to explain things gives you a kind of power" that is of paramount importance here (Bleicher, 2017, Show and Tell section). Doing so for an AI system would validate the power of the human over the machine, eliminate fear and instil trust in the system.

In the article "The Misleading Power of Internet Metaphors", the use of terms such as "cloud", "internet of things (IoT)", "smart" and "free" comes under scrutiny (Frischmann, 2018). Starting with the latter, the term "free" can be misleading as it refers solely to the monetary aspect of internet services. If one replaces "'free' with 'paid for with data' and 'possibly paid for with attention, labor, trust and even your mind'", the hidden content of the black box of the term begins to come to light (Frischmann, 2018, Free section). Transparency can also be increased if one attempts to pry into the opaqueness of the other terms mentioned. "Cloud" begins to make more sense when one talks about storing data in a complex network of various computers. However, using this term "served as an epistemological black box within which complexity was dumped and hidden" (Frischmann, 2018, Cloud section). In other words, the very use of the "cloud" metaphor creates a black box that makes the actual implementation of the internet cloud appear opaque to the naked eye. But, as already mentioned, "cloud" and "IoT" alike refer to an advanced use of already existing internet infrastructure. Nowadays, technology has increased interconnectivity at a global scale, giving the opportunity for new possibilities to arise. Using these terms creates a black box that obscures the actual processes involved. The same applies to the term "smart", a metaphor that builds on AI's black box. As stated, the term "appeals to our inclination to anthropomorphize tech" (Frischmann, 2018, Smart section). However, "the type of AI, how it works (or doesn't), who owns or controls it and many other details that vary tremendously across examples are hidden inside an epistemological black box" (Frischmann, 2018, Cloud section).

Additionally, the smartness mandate is radically criticized in the article. First, the term "conflates different forms of intelligence and makes it harder to evaluate differences in degree and kind" (Frischmann, 2018, Smart section). Despite the fact that "smart" can be very different depending on the technology, "smart seems unabashedly good, certainly better than dumb" (Frischmann, 2018 Smart section). But appearances can be deceiving. Sometimes, dumb technologies (such as cash) are very useful. Being "smart" is always contingent on the technology, the people involved and the context. Thus, the smart/dumb "dichotomy is itself pretty dumb" (Frischmann, 2018, Smart section). In short, the mandate to be "smart" may go hand in hand with AI's potential, but it is controversial, to say the least.

Another article from *Scientific American* deals with the issue of algorithmic bias (Young, 2020, p. 215). Namely, it deals with using the wrong algorithm to reach a specific conclusion. For instance, an algorithm created to assess the costs of healthcare may be used to infer the severity of an illness, thus providing wrong results (Young, 2020). Again, the issue of trust in AI comes into the forefront. It is proposed that opening the algorithm to more users by increasing inclusivity and transparency could be one of the solutions, along with constant testing of the algorithm for bias and discrimination against specific users or groups (Young, 2020, p. 215). In a similar vein, talking about AI accountability brings to light various and urgent societal problems, especially in the case of self-driving cars and the use of AI in courts of law. The article "Intelligent to a Fault: When AI Screws Up, You Might Still Be to Blame" specifically mentions these issues, raising the problem of opaqueness related to certain actors, such as the judge who is supposed to reach a decision based on an algorithm the workings of which are utterly covered in a veil of mystery (Greenemeier, 2018). Moreover, according to the article, depending on how society perceives AI and its usefulness (or, one should say, depending on how the ethics of AI are constructed) policy and governance are going to be shaped accordingly (Greenemeier, 2018). This conclusion points to the sociotechnical co-construction of AI in reference to its opaqueness and the transparency mandate. An article from 2016 points to the human biases that permeate AI applications (Emspak, 2016). As in humans so in AI machines, biases cannot be eliminated, and considering a machine neutral only increases the opaqueness of its black box (Emspak, 2016). Being cognizant of the biases that are embedded in data gathering and the design of AI machines is crucial to increase accountability and trust.

In the article "The Machine That Would Predict the Future", the proposition of dropping all of the world's data into a black box in order to extrapolate predictions from it is analyzed (Weinberger, 2011). There are many problems related to such an idea, starting from the complexity of social phenomena to the unpredictability of individual behavior. There are no concrete laws to follow in order to reach specific results. The author goes on to talk about the lack of trust in such a black box and the absence of understanding related to its outputs, which would make it impossible to use them to inform policy and governance (Weinberger, 2011).

6. Primary Data Analysis

The above presentation of *Nature* and *Scientific American* articles has been produced in reference to the main theme of the thesis, black box and AI. In this section, we will address the research questions more concretely in order to analyze these articles through the prism of the black box metaphor. We begin with the first set of questions, which include how the black box is produced, how it can be unpacked and whether we should try to do so or not. In reference to these questions, another one should be addressed, which is "how can we build trust in AI and what is its connection to the black box problem?" Moreover, the interplay between the biological and the mechanical by means of the black box metaphor will be presented, along with the white box metaphor—that is, a more transparent box, which can be used to mirror black box procedures.

Using the chosen STS framework—that is, the black box metaphor—one can go on to analyze the articles presented above. Whether explicitly or not, the construction of AI's black box is mentioned in many of these articles. Beginning with those from *Nature*, the construction of the black box can have many manifestations. The many layers in a deep learning instance create a black box as the information is diffused, modeling the procedures of the brain. Additionally, huge amounts of data, as in the case of big data, that humans are unable to process are left to the computational power of the machine. Given the biases embedded in the data, a black box is constituted in the initial stages of the process and is exacerbated during the processes to reach a certain output. Thus, the human interpreter has not actually understood anything about the results; only the machine has (Castelvecchi, 2016, pp. 21-23). Moreover, the constitution of the black box is propelled through certain mindsets that promote a positive axiological approach to machine workings, such as the belief that AI algorithms are used only to help physicians in their work ("All eyes are on AI", 2018). However, they can actually hinder scientific practice.

Drawing on Pasquale's work regarding the reciprocal relationship between algorithms and society in the digital age, the co-production of the black box is evident. In this case, the black box of AI, primarily a technological black box, is reconstructed as a technoscientific black box within which society shapes the usage of technology and at the same time it is shaped by it. Even the political process (governance of AI, for instance) becomes itself a black box as it is based on obscure procedures. There is certainly a lot to be done in order to open the black box of society, a technosocially constructed one with various factors contributing to its robustness (Aftergood, 2015, pp. 435-436). One has to analyze the various actors contributing to its construction, along with the various parameters that come into play.

To put the construction of the black box into context, we should point to Woolgar and Latour's discussion of the "facts" and how their reception either opens or closes the black box of science. As pointed out in Possati's article, when a scientific "fact" is still open for discussion. In other words, Latour's thinking about scientific controversies, the black box of science is open and its procedures can be accounted for. On the contrary, when discussion ceases, the black box closes and the scientific fact is no longer accessible to scrutiny (Possati, 2020, pp. 11-12). In the case of artificial intelligence, it is evident that, when talking about an algorithm that is helpful or one too complex to infer its procedures but useful with meaningful results, the discussion about its inner workings is closed. Mostly, this happens because the perceived costs of opening this black box seemingly outweigh the benefits of such an endeavor. Again, the black box of AI is co-produced through a multitude of factors, including axiological beliefs.

Turning to the question of whether we should unpack the black box of AI, one could point to the urgency of the mandate to open it in biomedicine. Pointing to Castelvecchi's article, among others, relying on an AI algorithm in order to provide an accurate diagnosis may be cumbersome or even dangerous when experts are ignorant of the exact procedures that lead to the specific diagnosis. In the same article, another important aspect of the black box problem came from Clune's team in 2014, which used techniques "that could maximize the response of any neuron, not just the top-level ones" (Castelvecchi, 2016, p. 23). They found that the black box problem "might be worse than expected" (Castelvecchi, 2016, p. 23). More specifically, "neural networks are surprisingly easy to fool with images that to people look like random noise, or abstract geometric patterns" (Castelvecchi, 2016, p. 23). Addressing the "fooling" problem is urgent, yet difficult. No proposed solution has reached universal acceptance (Castelvecchi, 2016, p. 23). In the case of a self-driving car, for instance, the fooling problem could have devastating results. This makes it even more important to actually try to pry open the black box of AI.

Additionally, most of the articles presented above made explicit reference to the necessity to open AI's black box in order to increase transparency and accountability regarding safety issues related to the use of AI algorithms. We will come back to the issue of trust in AI, which is obviously closely related to the mandate to open the black box.

Some ways to unpack the black box of AI are described in the aforementioned articles. One way could be reverse-engineering the procedures of the deep learning algorithms. Deep Dream tried to do so by locating the dependent variables responsible for the output given by the machine. Starting from the endpoint to reach the initial circumstances of the black box constitution, though not exactly reaching all the way back, we find the mandate to increase inclusivity by giving more users access to the algorithm. This way, people will be able to experiment with the algorithm at a high level, pointing to any biases and misgivings it may have. It could be said that this is a citizen science approach as, with basic training, AI machine's will be accessible to many people, blurring the strict boundaries between experts and lay-people. Eric Siegel's approach would be relevant here, as he asks for an increase in inclusivity in order to increase the transparency of machine and deep learning's algorithmic processes (Siegel, 2016).

Moreover, identifying the causal chain that leads from a specific cause to a specific outcome is of paramount importance in opening the black box of AI. It could be done by tracing the exact numbers that correspond to each knob and identifying the exact nexus or nexuses that create the causal chain. However, this is too demanding and even impossible in the case of deep learning. Moreover, the information is diffused into various knobs. Thus, identifying the numbers corresponding to each knob may mean nothing for the actual reconstruction of the information. For these reasons, experts proposed the use of another neural network, a more transparent one, in order to explain and interpret the workings of the target one. One of these is the hierarchical model presented in a *Nature* article, which works as a white box, the processes of which are mostly intelligible, in order to mirror AI's black box and pry it open (Xu et al., 2021, pp. 1-10).

Related to the mandate to open the black box of AI and, in any case, relevant to the discussion about black box's impact is the issue of trust in AI. Many articles point to the necessity to build trust and how the lack of it interferes with the actual application of AI. Starting from statistical analyses in which AI proved to be more accurate than humans, experts were hesitant to trust the results (Haiman, 2019, pp. 18-19). Again, in the case of biomedicine, trust in AI is a huge issue for experts and society alike (Burgess, 2019, pp. 374-375). In the manufacturing industry, as presented in one of the articles above, a way should be found to earn human trust in AI; to understand why people trust human experts more than AI, even if the experts are wrong (Li et al., 2021, p. 13564).

The complexity of the black box of deep learning algorithms makes it hard even for its creators to understand how they work. However, for the user to trust AI, she must first be able to understand it and predict its behaviour. The complexity that arises makes trust in AI very difficult because people must "depend on other superficial cues to make trust decisions" (Li et al., 2021, p. 13564). Trying to categorize them at the level of the individual, such cues "may include anthropomorphism, voice consistency, relationship type, and timeliness responding to AI" (Li et al., 2021, p. 13564). At an organizational level, trust in AI "is subject to cues from the institutional environment" (Li et al., 2021, p. 13564). The article points here to institutional theory, according to which "organizational and individual behaviour are influenced by regulative, cognitive and normative institutional dimensions" (Li et al., 2021, p. 13564). Since AI systems are usually introduced by managers and promoted by key promoters, attitudes from top managers, group leaders and AI promoters should have some influence on users' trust in AI, which suffices, according to the authors, for the conclusion that institutional theory is the appropriate methodological tool to use in order to analyse trust in AI (Li et al., 2021, p. 13564). From an STS perspective, the multiplicity of factors related to trust in AI actually makes institutional theory a very important, yet not the only appropriate, tool to view trust in AI systems.

As manifested in the articles presented in the previous section, trust in AI is a huge issue. Both those from *Nature* and *Scientific American* raise the problem of trust and propose ways to increase it in order to move on with the application of AI. From self-driving cars to the imaginary black box of science fiction, where all data could be stored in order to predict the future, trust in AI systems is closely linked to the transparency mandate. AI's black box, as long as it remains a black box, hinders trust and, thus, should be opened. Drawing on the article about the machine that could predict the future from *Scientific American*, it is important to stress that even if we had the computational power to create a black box such as "Rehoboam" from the popular TV show *Westworld*, the actual creation would be a total black box from which "knowledge" might come, albeit without understanding. That would eventually call for a redefinition of knowledge along these lines, but such a thing is unacceptable. Even without considering Gettier's counterexamples, which call into question our definition of knowledge, the outputs of such a machine would not even satisfy the basic definition of knowledge as "justified true belief" (Ichikawa & Steup, 2018). How can such knowledge be justified without resorting to blindingly trusting the machine? Finally, trust is inextricably linked to AI's black box, its impact on society and the co-production of AI as a socio-technoscientific artifact.

Moving to the last question about the connection between the biological and the mechanical through the lens of the black box metaphor's penetration into these domains, it is of great significance to mention that black boxes and white boxes are used interchangeably, jumping from one domain to the other. There is an interesting analogy between the brain and deep learning AI and, more broadly, between the biological and the mechanical, which is implied in Castelvecchi's article. According to it, neural "networks are also as opaque as the brain" because they do not actually store "what they have learned in a neat block of digital memory", but "they diffuse the information in a way that is exceedingly difficult to decipher" (Castelvecchi, 2016, p. 21). Maybe saying that they are "as opaque as" the brain is an overstatement, yet the analogy with the biological archetype of the black box and the fact that the processes within it are opaque present an interesting view of how the black box of AI is constructed and how we tend to perceive it.

Neural networks mirror brain processes, and, in doing so, they mirror the black box of the brain. For instance, how some synapses become stronger whereas others weaken represents a mystery that needs to be investigated. Information is not stored in a specific part of the network but, as in the case of the brain, it is largely diffused, making it harder to pinpoint its exact location and gain the complete picture of the causal chain that leads from the stimuli (or the cause) to the behavior (or the result). Additionally, in the case of stem cell reprogramming one encounters a situation where the black box of biology and ways to open it are mirrored to the black box of the machine: deep learning. In deep learning, stochastic gradient descent is used to accurately describe the probability distribution of various vectors, thus accounting for the randomness of the process. Using the black box metaphor to present both situations may not be the only way to go. For instance, the use of a stochastic model and its results in either domain can be utilized as a white box to penetrate the black box of the other, from biology to the machine or vice versa (Cyranoski, 2014, pp. 162-164).

Finally, the presentation and the discussion of the articles through the STS framework of this thesis shows how our research questions are manifested in relevant articles from popular science. The black box metaphor and the technoscientific and societal co-production of AI's black box, issues of transparency, trust in AI, and the mutual penetration of black box and white box models through the fields of the biological and the mechanical are pervasive in the contemporary literature about AI. Of course, the articles analyzed in this project can only present a small fraction of the literature, which is bound to increase as the interest in AI applications also increases.

7. Conclusion and Suggestions for Further Research

In conclusion, this project aimed at presenting the machine black box in the digital era; in the era of AI. Starting from an overview of the relevant secondary literature and mainly building a schema in which AI should be viewed, some important research questions regarding AI’s black box were raised. The analysis of the primary data (articles from the scientific journals *Nature* and *Scientific American*) was done through the prism of these questions.

Using the black box metaphor as a tool to assess AI is not only relevant but necessary in order to understand scientific endeavor in its entirety—i.e., as technoscientific and social. This methodological tool from the interdisciplinary field of STS was chosen due to its simplicity and its theoretical, descriptive, and exegetical value. More specifically, viewing AI applications through the lens of the black box metaphor helps our understanding of how opaque some, if not all, of its procedures are for certain stakeholders. For example, some deep learning machines can become so opaque that even their creators are at a loss regarding the exact processing of an input that leads to a specific outcome. Without delving much into the technical details of how algorithms and the various layers of a deep learning neural network actually work, our study of articles from popular science revealed the nuances of the black box rhetoric as far as the opaqueness of AI applications and their impact on society are concerned.

Throughout the whole thesis, the construction of AI’s black box has been depicted as a co-production of various technoscientific and societal factors. The constitution of the black box is neither one-sided nor only manifested by the complexity of the technology embedded in it. It is a process that involves how this complexity is perceived along with the choices society makes in order to increase or diminish the opaqueness of AI. As analyzed in detail in Pasquale’s book (Pasquale, 2015), algorithms shape and are shaped by society in a reciprocal, co-productive process. Of course, some of the actors involved may carry different weights into this process. A highly complex deep learning instance that provides socially desirable results may be the main component in the co-production of the black box, as a cost–benefit analysis would show that the benefits of opening this black box do not outweigh the costs of the process.

However, this is where the ethics of AI come into play. STS should be more concerned with the literature on ethics regarding AI's black box. The mandate to open the black box—i.e., the answer to our question of "why we should open the black box of AI?"—is closely linked to the ethics of AI or how society perceives them. In other words, a strict cost–benefit analysis—a narrow application of utilitarian ethics—is a societal choice regarding which ethical values it chooses to project on AI applications. In other words, we should open the black box of AI in order to deal with safety issues and help scientists in their work, but we should also open it to create a society where accountability and fairness are its basic tenants; an ethical society.

Most of the ways to unpack the black box of AI presented in this thesis had to do with its technical sides: searching for every knob and its correlation to the information, creating white box models to open the black box or including more users to test the outputs and the reverse-engineering. Of course, these are of primary importance when it comes to the modern AI black box. However, the issue of societal or policy choices and their implications for AI's black box were raised along with ways to understand and account for choices that obscuring instead of elucidating the black box. For instance, using the wrong algorithm for our goal is an issue that should be accounted for at a not-so-technical level. For all intents and purposes, the construction of AI's black box should be first understood in its multiplicity—i.e., as a co-production of many technoscientific and social factors. Then, the process of unpacking should become feasible, yet difficult in many cases.

Many of the articles presented in the primary data reflected on trust in AI. Probably, trust is an issue at the fringes of STS but a very closely linked one to the black box of AI and the mandate to open it. Increasing trust in AI by fostering procedural transparency, explainability and accountability was of paramount importance in the literature studied. It is proposed that, by doing so, it becomes feasible for the implementation of AI applications to expand in society without instilling fear and hesitancy to the end users. From experts to lay people, trust in AI is a critical factor that decides which AI application will be used or not. It goes without saying that the opaquer the AI black box the less trust in AI is manifested.

As far as the relationship between the biological and the mechanical, via the black box metaphor, is concerned, the present thesis found that biological and mechanical black boxes can be used to elucidate the procedures of each other. More specifically, the black box archetype, the human brain, is mirrored on deep neural networks, which, in turn, can be used as white boxes to get a glimpse on brain processes. On the other hand, biological processes, such as the reprogramming of stem cells, can be used as white boxes for AI machines. In this manner, AI machines present a peculiar type of machines as they try to mirror organic procedures. To follow Canguilhem's way of thinking, probably a mechanistic approach to their processes does not provide us with the complete picture. The direction of our train of thought should be from the biological to the mechanical, thus overcoming mechanistic frameworks that may limit our capacity to understand. However difficult to implement, this idea is more relevant than ever when it comes to AI as it is the epitome of the biological domain's expansion into the mechanical domain.

From the above discussion, it is quite apparent that the black box of AI presents a pressing issue that should be addressed thoroughly and carefully. Adopting an STS perspective, as in this thesis, raises some of the main concerns regarding the black box and shows how complex this black box is. As an interdisciplinary field, STS has various methodological tools to offer in order to critically reflect on and analyze AI's black box. Further studies should be made to understand it in all of its complexity and account for its sociotechnical impact. However, STS, stemming from a largely sociological tradition, tends, in the case of AI, to adopt a view focusing more on the societal aspect of technology. Of course, the technoscientific analysis of AI artifacts, how they are embedded in society and how they shape society are all very important points of focus for STS. But we would propose further research along two seemingly opposite sides. On the one hand, studies on AI and its black box should delve more into the technical domain of how algorithms can be designed and implemented in order to become more transparent, easy to use and accurate. On the other hand, these studies should meet with philosophy and the domain of ethics. The ethics of AI and big data represent a rapidly growing field, which is not that different from STS, nor should it be regarded as having a separate agenda. On the contrary, as mentioned in this thesis, issues of trust in AI (ethical issues) are at the core of the black box problem and the necessity to deal with it.

References

Primary

Aftergood, S. (2015). Big data: Stealth control. In: *Nature 517*, 435–436. https://doi.org/10.1038/517435a

“All eyes are on AI”. (2018). In: *Nature Biomedical Engineering 2*, 139–139. https://doi.org/10.1038/s41551-018-0213-2

Bleicher, A. (2017). Demystifying the Black Box That Is AI. In: *Scientific American*. https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/

Building natural trust in artificial intelligence. (n.d.). Retrieved June 18, 2021, from https://www.nature.com/articles/d42473-020-00352-0

Burgess, D.J. (2019). Illuminating the dark side of machine learning. In: *Nature Reviews Genetics 20*, 374–375. https://doi.org/10.1038/s41576-019-0140-4

Castelvecchi, D. (2016). Can we open the black box of AI? In: *Nature News 538*, 20. https://doi.org/10.1038/538020a

Castelvecchi, D. (2019). AI pioneer: ‘The dangers of abuse are very real.’ In: *Nature*. https://doi.org/10.1038/d41586-019-00505-2

Cyranoski, D. (2014). Stem cells: The black box of reprogramming. In: *Nature* 516, 162–164. https://doi.org/10.1038/516162a

Emspak, J. (2016). How a Machine Learns Prejudice. In: *Scientific American*. https://www.scientificamerican.com/article/how-a-machine-learns-prejudice/

Frischmann, B. (2018). The Misleading Power of Internet Metaphors. In: *Scientific American*. https://blogs.scientificamerican.com/observations/the-misleading-power-of-internet-metaphors/

Greenemeier, L. (2018). Intelligent to a Fault: When AI Screws Up, You Might Still Be to Blame. In: *Scientific American*. https://www.scientificamerican.com/article/intelligent-to-a-fault-when-ai-screws-up-you-might-still-be-to-blame1/

Haiman, Z. (2019). Learning from the machine. In: *Nature Astronomy 3*, 18–19. https://doi.org/10.1038/s41550-018-0623-9

Li, J., Zhou, Y., Yao, J. & Liu, X. (2021). An empirical investigation of trust in AI in a Chinese petrochemical enterprise based on institutional theory. In: *Scientific*

Reports 11, 13564. https://doi.org/10.1038/s41598-021-92904-7

Lo Piano, S. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. In: *Humanities and Social Sciences Communications 7*, 1–7. https://doi.org/10.1057/s41599-020-0501-9

Possati, L.M. (2020). Algorithmic unconscious: why psychoanalysis helps in understanding AI. In: *Palgrave Communications 6*, 1–13. https://doi.org/10.1057/s41599-020-0445-0

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. In: *Nature Machine Intelligence 1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.-G., Mahlein, A.-K. & Kersting, K. (2020). Making deep neural networks right for the right scientific reasons by interacting with their explanations. In: *Nature Machine Intelligence 2*, 476–486. https://doi.org/10.1038/s42256-020-0212-3

Weinberger, D. (2011). The Machine That Would Predict the Future. In: *Scientific American*. https://www.scientificamerican.com/article/the-machine-that-would-predict/

Xu, Y., Hu, M., Liu, H., Yang, H., Wang, H., Lu, S., Liang, T., Li, X., Xu, M., Li, Liu, Li, H., Ji, X., Wang, Z., Li, Li, Weinreb, R.N. & Wang, N. (2021). A hierarchical deep learning approach with transparency and interpretability based on small samples for glaucoma diagnosis. In: *npj Digital Medicine 4*, 1–11. https://doi.org/10.1038/s41746-021-00417-4

Young, T., N. (2020). I Know Some Algorithms Are Biased—because I Created One. In: *Scientific American*. https://blogs.scientificamerican.com/voices/i-know-some-algorithms-are-biased-because-i-created-one/

Secondary

Aizenberg, I., Aizenberg, N.N. & Vandewalle, J.P.L. (2000). Multi-Valued and Universal Binary Neurons: Theory, Learning and Applications. In: *Springer US*. https://doi.org/10.1007/978-1-4757-3115-6

Alaniz, R.J. (2020). Before the "Black Box": The Inputs and Outputs of Nineteenth-century Deep-sea Science. In: *Science, Technology, & Human Values 45*, 596–

617. https://doi.org/10.1177/0162243919881201

Big Data: What it is and why it matters. n.d. Retrieved June 13, 2021, from https://www.sas.com/en_us/insights/big-data/what-is-big-data.html

Bijker, W.E. (2015). Technology, Social Construction of. In: Wright, J.D. (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*. Elsevier, Oxford, pp. 135–140. https://doi.org/10.1016/B978-0-08-097086-8.85038-2

Bottou, L. (2010). Large-Scale Machine Learning with Stochastic Gradient Descent. In: Lechevallier, Y., Saporta, G. (Eds.), *Proceedings of COMPSTAT'2010. Physica-Verlag HD*, Heidelberg, pp. 177–186. https://doi.org/10.1007/978-3-7908-2604-3_16

Canguilhem, G. (2008). *Knowledge of Life*. Fordham Univ Press.

Chen, X., J. (2016). The Evolution of Computing: AlphaGo. In: *Computing in Science Engineering 18*, no. 4. https://doi.org/10.1109/MCSE.2016.74

Dechter, R. (1986). Learning while searching in constraint-satisfaction-problems. In: *Proceedings of the Fifth AAAI National Conference on Artificial Intelligence, AAAI'86*. AAAI Press, Philadelphia, Pennsylvania, pp. 178–183.

Engber, D. (2014). Who Made That Black Box? The New York Times.

Fahrenkamp-Uppenbrink, J. (2019). What good is a black box? In: *Science 364*, 38. 16-40. https://doi.org/10.1126/science.364.6435.38-p

Felt, U. (2017). *The handbook of science and technology studies*. Fourth edition. (ed.) The MIT Press, Cambridge, Massachusetts.

Fukushima, K. (1979). Neural network model for a mechanism of pattern recognition unaffected by shift in position - Neocognitron. In: *Transactions IECE*, J62-A (10):658–665.

Glikson, E. &Woolley, A.W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. In: *ANNALS 14*, 627–660. https://doi.org/10.5465/annals.2018.0057

Halpern, O., Mitchell, R. & Geoghegan, B. (2017). The Smartness Mandate: Notes toward a Critique. In: *Grey Room 68*, 106–129. https://doi.org/10.1162/GREY_a_00221

Hanley, R. (1997). *Is Data Human?: The Metaphysics of Star Trek*. Basic Books.

Hao, K. (2019). The computing power needed to train AI is now rising seven times faster than ever before, In: *MIT Technology*

Review.https://www.technologyreview.com/2019/11/11/132004/the-computing-power-needed-to-train-ai-is-now-rising-seven-times-faster-than-ever-before/

Hinton, G.E., Osindero, S. & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. In: *Neural Computation 18*, 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

Hwang, T. (2018). Computational Power and the Social Impact of Artificial Intelligence. In: *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3147971

Ichikawa, J.J. & Steup, M. (2018). The Analysis of Knowledge. In: Zalta, E.N. (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab. Stanford University.

Ivakhnenko, A., G. (1971). Polynomial theory of complex systems. In: *IEEE Transactions on Systems*. Man and Cybernetics. (4): 364–378.

Jensen, A. R. (1980). Chronometric analysis of intelligence. In: *Journal of Social and Biological Structures 3*(2). 103–122. https://doi.org/10.1016/0140-1750(80)90003-2

Kimura, A.H. & Kinchy, A. (2016). Citizen Science: Probing the Virtues and Contexts of Participatory Research. In: *Engaging Science, Technology, and Society 2*. 331–361. https://doi.org/10.17351/ests2016.99

Latour, B. (1987). *Science in action: how to follow scientists and engineers through society*. Harvard University Press; Cambridge, Mass.

Latour, B. (1999). *Pandora's hope: essays on the reality of science studies*. Harvard University Press; Cambridge, Mass.

Latour, B. & Woolgar, S. (1986). *Laboratory life: the construction of scientific facts*. Princeton University Press; Princeton, N.J.

Leslie, D. (2020). Understanding bias in facial recognition technologies: an explainer. In: *The Alan Turing Institute*. https://doi.org/10.5281/zenodo.4050457

Linnainmaa, S. (1970). The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. *Master's thesis*. Univ. Helsinki.

Marcheselli, V. (2020). The Shadow Biosphere Hypothesis: Non-knowledge in Emerging Disciplines. In: *Science, Technology, & Human Values 45*, 636–658. https://doi.org/10.1177/0162243919881207

Mascarenhas, M. (2018). White Space and Dark Matter: Prying Open the Black Box of STS. In: *Science, Technology, & Human Values 43*. 151–170. https://doi.org/10.1177/0162243918754672

McCallum, A. (2005). Information Extraction: Distilling structured data from unstructured text. In: *Queue 3*. 48–57. https://doi.org/10.1145/1105664.1105679

McCarthy, J. (2007). What is artificial intelligence? In: *Computer Science Department*. Stanford University, http://jmc.stanford.edu/articles/whatisai/whatisai.pdf

McCarthy, J. et al. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. http://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

Mukerji, C. (2001). Science, Social Organization of. In: Smelser, N.J., Baltes, P.B. (Eds.). *International Encyclopedia of the Social & Behavioral Sciences*. Pergamon. Oxford. pp. 13687–13691. https://doi.org/10.1016/B0-08-043076-7/03184-3

Nagel, T. (1974). What Is It Like to Be a Bat? In: *The Philosophical Review 83*. 435. https://doi.org/10.2307/2183914

Nilsson, N.J. (2010). *The quest for artificial intelligence: a history of ideas and achievements*. Cambridge University Press, Cambridge; New York.

Pacheco, P.S. (2011). Chapter 1 - Why Parallel Computing? In: Pacheco, P.S. (Ed.), *An Introduction to Parallel Programming*. Morgan Kaufmann, Boston. pp. 1–14. https://doi.org/10.1016/B978-0-12-374260-5.00001-4

Pandolfini, B. (1997). *Kasparov and Deep Blue, The historic chess match between man and machine*, Fireside Publications; New York.

Pasquale, F. (2015). *The Black box society: the secret algorithms that control money and information*, First Harvard University Press paperback edition. Harvard University Press, Cambridge; Massachusetts London, England.

Petrick, E.R. (2020). Building the Black Box: Cyberneticians and Complex Systems. In: *Science, Technology, & Human Values 45*. 575–595. https://doi.org/10.1177/0162243919881212

Pinch, T. (2015). Scientific Controversies. In: Wright, J.D. (Ed.), *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*. Elsevier, Oxford. pp. 281–286. https://doi.org/10.1016/B978-0-08-097086-8.85043-6

Raymond, A.H., Young, E.A.S. & Shackelford, S.J. (2017). Building a Better Hal 9000: Algorithms, the Market, and the Need to Prevent the Engraining of Bias. *Northwestern Journal of Technology and Intellectual Property 15*. 215.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning internal representations by error propagation. In: Rumelhart, D. E. and McClelland, J. L., (ed.), *Parallel Distributed Processing*. volume 1. pp. 318–362. MIT Press.

Rushton J., P. & Jensen, R., A. (2010). Race and IQ: A Theory-Based Review of the Research in Richard Nisbett's Intelligence and How to Get It. In: *The Open Psychology Journal, 3*. https://benthamopen.com/ABSTRACT/TOPSYJ-3-9

Russell, S. & Norvig, P. (1995). *Artificial intelligence: a modern approach*. Practice Hall. New Jersey.

Sanders, N. R. (2014). *Big Data Driven Supply Chain Management: A Framework for Implementing Analytics and Turning Information Into Intelligence 1st ed*., Pearson Education Limited.

Searle, J.R. (1980). Minds, brains, and programs. In: *Behavioral and Brain Sciences 3*. 417–424. https://doi.org/10.1017/S0140525X00005756

Shindell, M. (2020). Outlining the Black Box: An Introduction to Four Papers. In: *Science, Technology, & Human Values 45*. 567–574. https://doi.org/10.1177/0162243919883414

Siegel, E. (2016). *Predictive analytics: the power to predict who will click, buy, lie, or die*, Revised and Updated Edition. Wiley; Hoboken.

Smart, J.J.C. (2017). The Mind/Brain Identity Theory. In: Zalta, E.N. (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab. Stanford University.

Steen, M. (2015). Upon Opening the Black Box and Finding It Full: Exploring the Ethics in Design Practices. In: *Science, Technology, & Human Values 40*. 389–420.

Turing, A., M. (1950). Computing machinery and intelligence. *Mind LIX*. no. 236. https://doi.org/10.1093/mind/LIX.236.433

Weizenbaum, J. (1966). Computational linguistics. *Communications of the ACM: Computational Linguistics* 9. no. 1. Massachusetts Institute of Technology. Cambridge.

"What Is Big Data?" Oracle. n.d. Retrieved June 18, 2021, from: https://www.oracle.com/big-data/what-is-big-data/

Wiener, N. (2000). *Cybernetics or control and communication in the animal and the machine*. 2. ed. 10. print. ed. MIT Press; Cambridge, Mass.

Winner, L. (1993). Upon Opening the Black Box and Finding It Empty: Social Constructivism and the Philosophy of Technology. In: *Science, Technology, & Human Values 18*. 362–378.

Winston H. P. (1993). *Artificial Intelligence*. (3rd ed.) Addison-Wesley Publishing Company; USA.

Wylie, C. (2019). Glass-boxing Science: Laboratory Work on Display in Museums. In: *Science, Technology, & Human Values 45*. https://doi.org/10.1177/0162243919871101