UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ALGORITHMIC OPACITY : A NARRATIVE REVUE

DISSERTATION

PRESENTED

AS PARTIAL REQUIREMENT

FOR THE MASTER IN SCIENCE, TECHNOLOGY AND SOCIETY

BY

OLEG LITVINSKI

SEPTEMBER 2018

UNIVERSITÉ DU QUÉBEC À MONTRÉAL

L'OPACITÉ ALGORITHMIQUE : UNE REVUE NARRATIVE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN SCIENCE, TECHNOLOGIE ET SOCIÉTÉ

PAR

OLEG LITVINSKI

SEPTEMBRE 2018

ACKNOWLEDGEMENTS

A research project is not a solitary endeavour. This work would not have seen the light of day without the support, assistance and encouragement of many people.

I am grateful to my supervisor, F. Javier Olleros, professo at the Département de management et technologie, École des Sciences de la Gestion, Université du Québec à Montréal, for fruitful discussions, gentle guidance and unshakable assistance.

I am also grateful to researchers and personnel of the Centre interuniversitaire de recherche sur la science et la technologie (CIRST), Université du Québec à Montréal, for a favorable work atmosphere and many stimulating conversations.

Many thanks to my fellow students for sharing with me their research experience.

Finally, without my parents' support, I would never have been able to complete this project.

CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

ACM –Association for Computing Machinery

AGI – Artificial general intelligence

AI – Artificial intelligence

ML – Machine learning

MOS - Management and organizational studies

NGO – Non-governmental organization

ICT – Information and communication technology

IS – Information system

RÉSUMÉ

Dans un contexte de transformation numérique, l'usage répandu de systèmes algorithmiques est pris pour acquis dans une société moderne. Malgré leur popularité croissante, certains de ces systèmes présentent souvent un comportement opaque et imprévisible qui résulte dans un manque de compréhension et d'entendement de leur fonctionnement interne par les humains (le soi-disant 'problème de la boîte noire').

En visant à examiner l'opacité algorithmique dans le cas de modèles d'apprentissages profonds, cette recherche qualitative présente une revue de la littérature suivie du développement de concepts pour proposer une explication de l'opacité algorithmique en termes sociologiques.

Je définis l'opacité algorithmique comme le manque de compréhension par des humains de la logique des algorithmes ainsi que des résultats de leur exécution par une machine. En adoptant une perspective sociologique, le système analysé implique : (a) la présence d'acteurs sociaux, de leurs propriétés et de leurs actions; (b) leurs processus de prise de décision; (c) l'interprétation des résultats d'une exécution d'un algorithme par une machine; et (d) leur contexte social spécifique. L'opacité algorithmique peut être comprise alors comme une accumulation de compromis et d'autres dérivés de la prise de décisions humaines pendant le cycle de vie de systèmes algorithmiques.

Ces résultats pourraient attirer surtout l'attention des chercheurs en gestion et design organisationnel, ainsi que d'autres chercheurs en sciences humaines et sociales. Mais ils pourraient être contestés par leurs collègues de technologies d'information et la communication, au niveau de l'évaluation de qualité.

Divisé en deux parties (la création de concepts et leur évaluation empirique), mon futur travail consistera à plusieurs études de cas ayant pour but d'identifier les causes de l'opacité algorithmique et de trouver des façons de traiter les répercussions de cette opacité dans les divers contextes organisationnels.

Mots clés: algorithme, opacité algorithmique, recherche qualitative, apprentissage machine, algorithme apprenant, apprentissage profond, entreprise.

ABSTRACT

In a context of digital transformation, the widespread use of algorithmic systems is a given in modern society. In spite of their growing popularity, some of these systems often exhibit a puzzling behaviour resulting in a lack of understanding and comprehension of their inner workings by humans (i.e. the so-called 'black-box problem').

Focused on algorithmic opacity in the case of deep machine learning models, this qualitative explanation-driven research takes the form of a literature review followed by conceptual developments aimed at explaining algorithmic opacity in sociological terms.

I define algorithmic opacity as the lack of human understanding of algorithmic logic and its execution by a machine. From a sociological perspective, the overall mechanism involves the presence of (a) social actors, their properties and actions; (b) their decision-making processes; (c) their interpretation of the results of an execution of an algorithm by a machine; and (d) their specific social context. Algorithmic opacity then may be understood as a result of the accumulation of trade-offs and other by-products of human decision-making during the life cycle of algorithmic systems.

These results may appeal mostly to management and organizational researchers and practitioners as well as other social scientists, but they might be challenged by ICT people working on quality appraisal and qualitative design.

Divided into a concept building and an empirical part, my own future work will consist of multiple case studies aiming to identify the causes and mitigate the effects of algorithmic opacity in various organizational contexts.

Keywords: algorithm, algorithmic opacity, qualitative research, enterprise, deep machine learning, learning algorithm.

INTRODUCTION

The word 'algorithm' is historically associated with a specific technological device, such as a personal computer in the late 20th century and a smartphone or tablet more recently. In both public and private spaces lately, this word has escaped its original definition. Once reserved for specific concepts and associated mostly with technological devices and practices, the algorithm become currently used for defining and describing various socially-related phenomena, such as team' or enterprise' networks and capabilities (Trexler, 2008; Lester, 2017).

As a necessary and intrinsic element of digital transformations (Olleros & Zhegu, 2016a; Floridi, 2009, 2014), algorithms affect human practices in various fields of activities and attract researchers from very diverse domains. For example, legal scholars and political scientists explore the ethical implications, as well as the accountability, transparency and auditing of such systems (Diakopoulos, 2016; Mittelstadt et al., 2016; Woolley & Howard, 2016). Communication researchers are interested in algorithms and their impact on human behaviour through social media (Bechmann, 2017; Eslami et al., 2016). Other social scientists consider algorithms as concealing, explicitly or implicitly, human actions and intentions, collective and personal values and norms (Bucher, 2017; Gillespie, 2017; Willson, 2017). More specifically, Bolin & Schwarz (2015) are interested in the role of algorithms in capturing implicit social structures. Moreover, Beer (2016) investigates the question of algorithmic systems and their influences in ordering social structures.

For management and organization-oriented researchers, the concept of algorithm occupies two closely related places, in big data analytics (Ambrose, 2015; Ekbia et al., 2015) and in deep machine learning (Armando, 2017; Dhar, 2016). In both cases, algorithmic systems are bounded by an established social institution such as an enterprise or an NGO (non-governmental organization).

Among the common themes across these streams of research, two are the most interesting for my own research. The first one includes transparency and accountability on the part of owners and managers of algorithmic systems (Diakopoulos, 2017; Ananny & Crawford, 2016; Danaher, 2016). Scholars in the domain of artificial intelligence and machine learning are also exploring characteristics such as algorithmic safety (Vassev, 2016) and robustness (Russell et al., 2015). The second theme concerns algorithmic opacity, the lack of human understanding of algorithmic logic and results, sometimes called the 'black-box' problem (Diakopoulos, 2014; Perel & Elkin-Koren, 2017; Tzeng & Ma, 2005). Once combined in a larger picture, both of these themes may be considerd as necessary elements of research efforts toward building artificial intelligence possessing human-level capabilities across a vast range of tasks and, moreover, explaning its own performance in such tasks (Goertzel, 2014; Müller & Bostrom, 2016).

One specific kind of algorithm, identified by the generic term of learning algorithm (Olleros & Zhegu, 2016b), deserves special attention as it modifies its structure as a consequence of a continuous interaction with the environment and data, all of which are somehow bounded by existing social structures, such as enterprises. This peculiar feature accentuates the above-mentioned opacity in the case of deep learning algorithms operating on big data (L'Heureux et al., 2017; Najafabadi et al., 2015).

On the other hand, learning algorithms are responsible for many recent successes in cognitive tasks based on perception, and in more complicated situations such as games requiring strategic decision-making and foresight, as well as cooperation and other kind

of intelligent behavior. Among the former, visual object recognition, as well as text and voice identification and generation are the most prevalent and promissing (Russakovsky et al., 2015). As for the latter, various strategic games (e.g., Go and poker) are the latest examples of machine learning attaining better-than-human-level performance quite rapidly (Moravčík et al., 2017; Silver et al., 2017). These achievements attract various groups of social actors (e.g. business community, elected officials, social activists, etc.) with the promise of solving problems or advancing performance to the point that international organizations and governments are starting to pay attention to this domain (Biegel & Kurose, 2016; National Research Council, 2013).

To sum up, the success and omnipresence of algorithmic systems in modern life, combined with some of their characteristics such as impenetrable and inexplicable inner workings, enigmatic and abstruse functioning, present an interesting phenomenon to investigate. We shall try to disentangle these problems and present a plausible account of the current mosaic of hypotheses, facts, models and theories about algorithmic opacity.

The highlights of this report are as follows. The evasive nature and long-standing issue of algorithmic opacity warrant a shift in perspective and the adoption of a process-oriented sociological view. The explanatory mechanism includes social entities, their decision-making processes and the interpretations of algorithmic logic and results in various social contexts. The rise of algorithmic opacity may be articulated as the accumulation of by-products of human decision-making during the algorithmic life-cycles in different situations. Further refinements of this account may provide guidelines, suggestions for mitigating the effects of algorithmic opacity and other digital phenomena exhibiting enigmatic behaviour.

The outline of this report is as follows. The first section, a conceptual background, presents a broad survey of algorithms, their opacity and related questions. This survey

will cover the definition of algorithm, the history of the idea of learning algorithms, artificial intelligence, an overview of deep learning models and architectures, as well as a brief overview of existing theories of learning algorithms. Moreover, this broad survey will also include a discussion of opacity and its close neighbors, and will present existing solutions. Lastly, the discussion will highlight more general trends in machine learning and artificial intelligence alongside broad remarks on the state of research into learning algorithms.

The second section, on objectives and methodology, details the research questions and methodological approach. This approach is a combination of a narrative literature review and a specific type of concept development serving together for articulating a coherent explanation.

The third section, developing an explanation, starts with a presentation of origins of algorithmic opacity and its overall mechanism in sociological terms. This chapter concludes with a subsection about the history of algorithmic opacity.

Finally, this report concludes with some implications of our research into algorithmic opacity (i.e., the transferability of results), the limitations of this research (i.e., possible objections) and avenues for further study (i.e., refinements of the mechanism and empirical validation).

CHAPTER I

CONCEPTUAL BACKGROUND

1.1 Historical landmarks

Before pursuing a general discussion of learning algorithms and their associated issues, it is important to anchor the topic in its historical context and review the ideas that have contributed to its evolution. Among these ideas, artificial intelligence (AI) is arguably the most important one: it includes learning algorithms in general and deep learning in particular.

In the history of AI, three main periods are worth mentioning as the ideas of learning, knowledge, agent and agency, and the social and natural world are all constantly present. The Dartmouth conference of 1955 is considered a turning point in the history of AI, as it propelled AI into the area of a serious research (McCarthy et al., 2006). After the conference, the logic and symbolic era continued until the connectionist turn of the mid-1980s. This era produced a panoply of expert systems that were meticulously constructed by humans trying to encode knowledge and practice in logically manipulated symbols. Despite being brittle and specialized in only a handful of closely related tasks, these systems gained popularity among large industrial players.

As part of the mid-1980s connectionist turn in the cognitive sciences, research into AI and machine learning (a branch of AI) adopted the neural networks metaphor (Mira, 2008). Borrowed from cognitive scientists who believed that a single region of the mammal brain is responsible for all perception, and combined with computer scientists'

essential principles and practices, this metaphor initially produced rather mixed results. In the late 1980s, researchers were aware that AI's great promises coexisted with a lack of understanding about the inner workings of artificial neural networks (Gallant, 1988). This lack of understanding – labeled "the black-box problem" – did not impede the gradual development and diffusion of artificial neural networks in other research domains and fields of practice, such as medicine (Barreto & de Azevedo, 1993).

In 2006, a team of researchers led by Geoffrey Hinton (Hinton et al., 2006) solved an enduring obstacle in AI performance: the "exploding and vanishing gradient" problem which is summarized as the premature abortion of algorithm' execution during mathematical operations (Bengio et al., 1994)[1]. This achievement marked the advent of a new era in the development and diffusion of neural networks under the name of deep neural networks. This new type of neural network constitutes the essence of deep learning architecture. Following the breakthrough in neural networks' performance came a proliferation of practical applications across a variety of industries, which gradually garnered public attention. However, the "black-box problem" is still haunting deep architecture networks, and increasing numbers of politicians, social actors and industrial players are calling for greater efforts aimed at solving this problem.

1.2 Algorithms

1.2.1 On the definition of 'algorithm'

[1] For example, the consecutive multiplication of large numbers will give an ever larger number while attaining a practical limit of resource availability in the machine; the same logic applies to the small numbers and other mathematical operations (Bengio et al., 1994).

As digital artifacts situated in a social realm, algorithms interest researchers from various fields and domains. These researchers employ different approaches and adopt multiple perspectives.

Scholars in computer science or ICT trace their definition of algorithms back to the work of Alan Turing in applied mathematics (Turing & Copeland, 2004). Two types of definition are worth mentioning. Among the examples of the first type, Gurevich (2012, 2013, 2015) tries to define algorithms based on the mathematical concept of "abstract state machines." Scholars investigating the second type derive their definitions from the basic mathematical notion of recursion. Moschovakis (1998, 2001) and Moschovakis et al. (2008) have pursued this type of definition. Finally, Yanofsky (2011) attempts to combine and enhance these two lines of inquiry in his own definition of algorithm.

Usually expressed as a group of axioms and theorems (Hoare, 1969) combined with a mathematical proof (i.e., a chain of operations constructed by following mathematical rules), these definitions of algorithm suffer from some major flaws from a social-science perspective. Firstly, they lack human-specific characteristics, attributes and properties, such as beliefs, intentions, actions and reflections. Secondly, their abstract characteristics make them difficult to use directly for the investigation of social phenomena. They are therefore not appropriate in situations where social interaction between individuals and groups involves human-specific constructs such as norms and values. Moreover, their application precludes the use of a proper level of investigation, as the boundaries are not clear and the separate concerns among the domains of research are difficult to delineate.

A definition proposed by Kowalski (1979) – i.e., "algorithm" as a combination of logic and control mechanisms – has recently gained popularity among communication and other social scientists (Roberge & Seyfert, 2016). Although simple, versatile and easy to remember, this definition seems to have been rejected by computer scientists in the

last few years. Thus, if used, it may add to the existing confusion, as researchers and practitioners from different fields typically work with completely different concepts.

In search of a definition of algorithm appropriate for management and organization studies scholars, this project follows Rapaport (2012, 2017) and especially Hill (2016). This approach makes it possible for human-specific attributes and notions to be included in the definition without the latter being restricted to a particular stream of research, domain or field of practice.

1.2.2 Defining 'algorithm'

Among recent papers addressing the definition of "algorithm," Hill's work (2016) deserves special attention. The definition given in that paper is as follows:

> "An algorithm is a finite, abstract, effective compound control structure, imperatively given, accomplishing a given purpose under given provisions." (Hill 2016, p. 47)

Three characteristics of Hill's definition make it appropriate for this project. Firstly, because it is formulated outside the hard sciences, it can be adapted to the social sciences and, consequently, to management and organizational studies. Secondly, it does not incorporate the researchers' hidden assumptions and beliefs. Thirdly, the pragmatic nature of this definition makes it appropriate for eventual use in various research' domains and fields of practice.

In this definition, three distinct elements are identifiable: context, purpose and description. The first, the context (i.e., "a given provision"), refers to an algorithm's environment, or the conditions and resources that are necessary and sufficient for it to be realized and executed. In the present project, "context" is understood not in the narrow sense of ICT or computer science (e.g., easily identifiable and numbered inputs,

variables and factors), but in a broader sense, encompassing everything that is not explicitly classified under the rubrics of purpose or description.

The second element of the definition, the purpose (i.e., "a given purpose") relates to the aim, goal or objective of the algorithm accomplishing a task. The word "task" expresses what the algorithm does from an external point of view. Moreover, this aim, according to Hill (2016), must be described in the active voice (i.e., "imperatively given"), because it implies some sort of human or other type of agency. In the present circumstances, humans define purpose in the form of intentions and expectations. The final element of the definition, a description involving "a finite, effective compound control structure," pertains to the algorithm's internal structures, components and mechanisms.

According to Hill (2016), the relationships between the algorithm's context, purpose and description are constrained by the fact that knowledge of the description is not sufficient to deduce the provision and the purpose (Hill, 2016, pp. 48). In other words, all the elements need to be clearly formulated before the algorithm can be executed or implemented by a machine.

The question of boundaries also acts as a constraint on the definition of an algorithm and, consequently, of algorithmic systems. In this research project, this constraint could take the form of a social structure. Enterprises and NGOs are examples of highly formalized social structures, while online communities united by a common interest (e.g., around outdoor sports activities) are examples of less formalized social structures.

The three elements described above – context, purpose and description – make it possible to clearly distinguish between internal and external perspectives in cases involving machines and humans. The latter perspective requires the explicit presence of a cognitive agent. Moreover, Hill's definition makes it possible to clearly separate

the concerns of social scientists and management researchers from those of computer scientists and ICT experts.

Two other concepts mentioned by Hill (2016, pp. 56-57) are necessary for further analysis of learning algorithms and their opacity: the machines' execution of algorithms and the human understanding of the results. The focus of this research project is on the results of algorithmic system execution, as perceived by human beings. An alternative to this view would require answering questions about what constitutes the inner workings of algorithmic systems. In the case of deep learning, it could include different types of neural networks: for example, deep belief networks, or recurrent neural networks.

1.2.3 On learning algorithms

Among the existing definitions of learning algorithms, one is worth closely examining as it is the most widely used in machine learning communities. From a technological perspective, learning algorithms are part of the domain of machine learning – an area of statistics emphasizing the use of computers for calculating mathematical functions . Goodfellow et al.'s recent book (2016, pp. 98-99) provides a useful starting point for discussion. Their definition includes three main components: a task (e.g., a classification), a performance measure for evaluation purposes and the experience (i.e., the learning process for the observer). There are two kinds of experience: one is supervised and the second is unsupervised according to the role and actions of the observer. In the first case, the observer guides the learning process in order to achieve a desirable goal. In the second case, the observer abstains from direct interaction with the learning algorithm (idem., pp. 102-103). As a side note, it is worth mentioning that reinforced learning differs slightly from both types of experience, as the observer offers incentives, but not necessarily guidance, to the algorithmic system (idem., p. 104).

Some deficiencies make it problematic to use this definition for the purposes of investigating social phenomena. Firstly, social entities are not explicitly present; they

are outside the learning algorithm. As such, human agents and their actions are not taken into account from the beginning. Secondly, their role and place in the learning process is greatly underappreciated: they are reduced to the representation. Other considerations, such as social, ethical and legal consequences, are relegated to other domains of research. Moreover, some empirical evidence points to the shortcomings of algorithmic systems built with this definition in mind. For example, Ching et al. (2017) note that real-world problems are very hard to map properly onto machine learning tasks.

1.2.4 Learning vs. traditional algorithms

A multifaceted concept, "learning algorithm" encompasses social and technological dimensions. From a MOS (management and organizational science) perspective, learning algorithms necessitate social actors. Indeed, social actors' attributes and actions continuously modify learning algorithms' structure and functioning. From an ICT perspective, data and programs take centre stage. Therefore, a combination of the MOS and ICT perspectives could make it possible to discover and explain trends and patterns that would otherwise be missed.

Learning algorithms blur some boundaries that traditional algorithms respect, especially as we move toward more autonomous forms of AI programming. For example, the separation between programmer and machine and that between software and data. Deep learning approaches also worsen the inherent lack of transparency of algorithmic systems (Goh et al., 2017; Liu et al., 2017; Yu et al., 2017). Table 1.1 provides a characterization of differences between traditional and learning algorithms.

Traditional and learning algorithms engender very different social and economic impacts in the real world. Being able to reach beyond the range of routine activities into more creative domains, learning algorithms will tend to be more disruptive of current best practices. Moreover, while the consequences of individual learning algorithms may be relatively foreseeable and manageable, the interplay between larger

Table 1.1 Comparison of traditional and learning algorithms*

		Traditional	Learning
Human	Agency	Predominant and mostly constant in time; any deviation is error or misbehaviour; human-coded rules define action	Ranges from minimum to pronounced; in the long run, most actions are without human direct involvement; a decrease of agency with time; difficult to attribute agency
	Activity	Initial development of a static structure; feature added manually; training phase is absent; constant human oversight	Initial development (i.e. overall objectives, results wanted); extensive and laborious training; difficulties of assessment and control; new expertise is required
	Aim	Quantifiable attributes (e.g. accuracy, precision);	A range of hardly quantifiable attributes (e.g., fairness) become important
Machine	Data	Mostly from tightly-controlled sources; highly structured (e.g. curated)	Not so highly structured; Mostly in the form of Big Data; unclear distinction between data and program
	Program	Minor human-crafted modification over time at a human scale and pace	Major modifications without human involvement at a rapid pace
Span		Confined to distinct units (i.e. social groups, enterprises)	Across many spheres, layers, and units which are not close by usual metrics
Example		Information systems separated by social barriers (e.g., rules and regulations): legacy desktop office suite; enterprise accounting systems, etc.	Mostly digital, large-scale and massive platforms crossing geographical, economic, social and other barriers (e.g., online search, weather forecast, new approaches to cybersecurity, etc.)

* "Span" refers to the breadth of impact on economic sectors, politics, culture, etc.; "Agency" refers to the autonomy and indeterminacy of possible actions.

clusters of complementary and competing algorithmic systems will be full of surprises.

1.2.5 Existing theories of learning algorithms

Three very different theories can help us to understand the inner workings of learning algorithms.

A Statistical Learning Theory was first proposed by Valiant (1984, 2013) in the context of applied mathematics. It was refined by Vapnik (2000), who gave it its popular name and has been widely accepted since the mid-1990s. This theory explains the functioning of machine learning in probabilistic terms. Inheriting optimisation properties from mathematics, this theory combines a manifold hypothesis (i.e., real-world data could be represented as a geometrical manifold, which, once dissected, could offer useful information about the feature of interest) that has proven to be useful in the task of classification, ranking and pattern recognition. In the early 2000s, probabilistic visual modelling encountered mixed success when it was applied in diverse contexts for exposing the mechanisms of machine learning. Olden & Jackson (2002) claimed to have solved the problem of neural networks' opaqueness, but the problem remains to this day. In short, Statistical Learning Theory might explain the inner workings of deep learning, but only partially and from a statistical standpoint.

As an alternative to the probabilistic model, some researchers have pursued other directions. Following the principles of information, as defined by Shannon (2001), Tishby et al. (2000, 2015) affirm that information in neural networks flows through defined places, as per a bottleneck principle, without losing sight of the feature that was of interest. From this perspective, generalization occurs through a noise reduction mechanism. Although promising, this theoretical proposal will require more time and effort as well as empirical evidence if it is to yield results.

Finally, some researchers have attempted to explain the workings of deep learning through the lens of physics. Lin et al. (2017) argue that principles that govern the natural world – for example, symmetry, locality and others – could give insight into

how deep learning works. One way to interpret this statement is to say that common principles of composition, hierarchy and layering are all hallmarks of deep learning.

As shown by these three theoretical options, a single approach to the functioning of deep learning will necessarily be limited. Deep learning needs to be examined from multiple perspectives, including social ones.

1.2.6 Deep learning : a branch of learning algorithms

A discussion of learning algorithms must include an overview of deep learning architectures, models and modes. More specifically, it is necessary to provide an overview of their characteristics in regard to human decision-making, as decision-making is one of the most fundamental human attributes.

The main idea underlying deep learning is that a single isolated unit is meaningless, but when it is combined with its neighbours, it gains some meaning. For example, in the case of visual perception, a standalone pixel means nothing to the human eye, but a combination of many pixels can disclose a lot more information. This basic principle also applies to sound and may even be relevant for other kinds of human perception (e.g., smell, taste and touch).

In this report, the terms "decision-making" and "decision-making process" refer to the process of choosing between at least two options or finding a consensus to resolve a dilemma about a desired future (Kerr & Tindale, 2004). This applies at both the individual and small-group levels. The process may be assisted by an artificial agent, such as an algorithm.

Table 1.2 provides a high-level overview of deep machine learning in regard to human decision making. Table 1.3 presents the strengths and weaknesses of deep learning architectures and modes for assisting a decision-making process. More in-depth overviews of deep machine learning are provided by Jordan & Mitchell (2015) and

LeCun et al. (2015). For a still more technical and detailed overview, the work of Schmidhuber (2015) and Arulkumaran et al. (2017) may serve as a starting point. From a multidisciplinary point of view, we have excellent surveys by Gawehn et al. (2016), Goh et al. (2017), and Ravì et al. (2017).

Table 1.2 Deep learning models and human decision making*

Model	Aims	Advantages	Disadvantages
Generative	Generate new or previously unknown traits and features	Consensus is attainable; contention is short-lived; decision making process is straightforward and undisputable; social norms and values are explicit; range of short-term prediction about immediate future state that may escape human attention; lower cognitive load on the human side	Domain and field specific difficulties (e.g. training, data quality, expert knowledge available); hidden biases; misjudgements are hard to detect; easily steered toward misbehaviour; hardly detectable persistent misbehaviour
Discriminative	Emphasizes existing traits and features	Eventual automation; easy to trust	Requires judgement and long-term experience

* Hybrid models are omitted as they combine both models' advantages and disadvantages.

Table 1.3 Deep learning architectures and modes*

Architecture & mode	Aims	Advantages	Disadvantages
Deep (implement connectionist paradigm of cognitive sciences)	Generate new options (i.e. generative) or separate existing (discriminative)	May have excellent performance for unitary or closely related features and behaviors to choose from; mostly acceptable for experts solutions	Major hurdles before usage; weak or impossible for human-specific reasoning; problematic long-term forecast; absent situational awareness at multiple scale; dubious cross-field application and cross-domain expertise;
Shallow (follows a symbolic paradigm)	Mostly recognition and identification of existing options	Easy to understand functioning and results; optimal solutions may be attainable	Domain and field specific; isolated cases in terms of narrow prediction; very short term and tailored prediction
Supervised and reinforced	A target is provided by external agent and reward is offered	Attaining a well-defined objective within a given time frame	If objective is unclear or changing, possibly wrong advise
Unsupervised	Only (raw) data is provided, without any human guidance	Detecting unexplored, previously unknown area	Time frame unclear; proposed actions ethical and legal considerations

* "Mode" refers to the process from an external point of view, while "architecture" refers to the internal structure of deep learning.

Among the common patterns found in Tables 1.2 and 1.3, two are most salient, namely connecting the past with the future and fields of use with human expertise. Generative models, unsupervised modes and deep architectures may be useful in assisting humans to explore possibilities and choose between courses of actions in view of achieving a desired goal. The combination of discriminative models, shallow architecture and

reinforced modes may provide an additional explanation for past mistakes, as it makes it possible to generate alternative scenarios of action leading to improved outcomes.

Although the distinctions in term of features and behaviors depend mostly on the specifics of the problem, the deep learning exhibits a plethora of shortcomings that are context and problem dependant. With regards to fields of applications, deep learning currently provides excellent results with tasks and activities when a perceptual evaluation of results (i.e., sound and vision) is realistic and attainable. Image and video recognition and generation, as well as hearing and speech recognition and creation, are the most prominent examples of deep learning's success. It is important to note, however, that this success has always been facilitated by human judgment and other high-level cognitive capacities (e.g., strategic planning, fine-grained coordination, mutually beneficial cooperation, etc.). Inherent to humans across cultures and history, those capacities are honed through prolonged practice and encoded in socially acceptable norms and values.

1.2.7 Opacity and related concepts

In addition to algorithmic systems and their opacity, other closely related concepts are worth discussing as a means of isolating the issue of opacity from possible overlaps, such as transparency, accountability, fairness and interpretability. Four recent contributions to the literature reflect the variety of views on opacity and related concepts. Table 1.4 summarizes the definitions of interpretability, transparency, accountability and fairness given in each document, with special attention to the way opacity features in each definition.

Although algorithmic opacity is mentioned in all cases, it is never confused with any other concept. Whereas the meaning and description of interpretability, transparency, accountability and fairness may differ according to a particular perspective and domains of research, the distinction between opacity and these four concepts is evident.

Table 1.4 Concepts similar to opacity

Authors	Transparency	Accountability	Fairness	Interpretability
Chakraborty et al., (2017)	Prerequisite for accountability; a desired feature of an algorithm	Ability to inspect model after execution; a desired behaviour; opaqueness is a feature of algorithmic process	Equality; lack of discrimination	Opposite of explainability; defers decision to an agent; agent interprets an explanation
Lepri et al., (2017)	Influence accountability; omission brings the inequality; opposite of opacity; explanation of a process leading to an easier decision		Lack of bias & inequity; types of fairness (i.e., individual and group); equality of opportunity	Opacity is the opposite of interpretability
Vedder & Naudts (2017)	Attribute of accountability (i.e., provides justification and explanation for actions or decisions)	Achieving accountability by diminishing algorithmic complexity		Interpretability is external to an algorithm; opacity is an attribute of algorithm
Binns (2017)		Accountability as providing reasons, justifications & explanation for decisions made; opacity is a separated attribute; opacity as unintelligible explanation		

Furthermore, all four concepts require human decision making and some form of social structure. Lastly, Vedder & Naudts (2017) examine the European Union's new data regulation on accountability and transparency for big data and algorithms. Writing from a legal perspective, the authors advocate for a more fine-grained oversight of the parties involved (e.g., individual, industrial actors, etc.). They call for a more comprehensive assessment of the societal impacts that come with the widespread diffusion of algorithms and argue that this assessment must also consider adjacent technologies such as big data.

1.2.8 Algorithmic opacity and proposed solutions

It is useful to continue our investigation into the issue of algorithmic opacity with a presentation of existing typologies as well as proposed solutions. Table 1.5 summarizes some recent contributions to the literature on opacity. Looking at this table, some observations are worth to highlight. First, the overlap of descriptive elements is prominent. For example, the words 'code-audit' and 'technical transparency' involve an expertise combined with practice and describe various facets of the core activities across disciplines. Furthermore, although authors may diverge in their opinions on the characteristics while proposing narrow-focused solutions, the initial appearance of algorithmic opacity might be overlooked. The short message from this table is the following: as a fluid and undefined concept, algorithmic opacity emerges from a social context and requires some digital artifacts.

As for the existing solutions to opacity, they may be divided by domain of research and field of practice. From a computer sciences and ICT perspective, researchers have been aware of the problem since the late eighties (Gallant, 1988). Earlier techniques included qualitative and quantitative methods and measures that were essentially linked with mathematics and statistics (Olden & Jackson, 2002). More recently, researchers have focused on visualization (Seifert et al., 2017) and enhanced interpretability and transparency of neural networks and machine learning (Chakraborty et al., 2017;

Miotto et al., 2017). Among these techniques, quantitative causal models and built-in interpretability (Datta et al., 2016) have been added to the potential solutions from previous decades. In spite of these ongoing efforts to enhance algorithms' transparency and interpretability, it would be difficult to imagine a characteristic labeled "opacity" attached to a specific algorithm in the works of ICT researchers.

Table 1.5 Typology of algorithmic opacity

Authors	Types and categories	Dealing with opacity
Burrell (2016)	Intentional: institutional actors introduce the opacity; technical illiteracy: possessing skill and knowledge for creating and operating an algorithm; operational: behaviour of (software) application	Auditing: code-audit; interpretability; avoidance of algorithms; simplification (e.g., 'feature extraction')
Geiger (2017)	Process of learning an institutional culture	Algorithmic literacy
Veale (2017)	Opacity as protective measure against internal and external treats (i.e., intentional abuse of existing routines)	Enhancing transparency
Robbins & Henschke (2017)	Technical: difficulty of comprehension of inner working of technology; algorithmic: properties of algorithm handling data; legal: obtaining data (and algorithms) from other countries	Transparency through disclosure of information; technical and algorithmic transparency

Working from a social science perspective, Zarsky (2016) argues that algorithms should be excluded in certain cases and advocates for more transparency and audit as well as indirect measures. For the authors cited in Table 1.5, transparency is an appealing solution. However, in an institutional context, the pursuit of transparency may be limited by what Stohl et al. (2016) have called the "transparency paradox".

According to them, three empirical attributes (i.e., availability, accessibility and approval) of information and data in organizational settings do not reflect a simple linear relationship between transparency and opacity. Through various mechanisms, the pursuit of greater transparency might end up increasing opacity. In their view, inadvertent opacity happens if humans are overwhelmed by information and data when they need to take an appropriate course of action[2]. According to them, a second path leading to opacity, strategic opacity, resides in intentional oversupply of information so that its receiver might miss the most important one. In short, their work undermines the assumption that algorithmic transparency always leads to an attenuation of opacity.

Without specifically targeting algorithmic opacity, Kitchin (2016) proposes a range of techniques that cover almost all of algorithmic systems' social aspects. Inspired by ethnographic studies, these techniques include: (a) investigation of human-produced codes; (b) reverse engineering of digital artefacts, including algorithms; (c) reflective production of such artefacts and algorithms; (d) scrutiny of the intentions and actions of teams producing artefacts and algorithms; (e) investigation of other contextual socio-technological elements; and (f) inquiry into the use of previously deployed algorithms. Together, these techniques offer a holistic approach to algorithmic opacity and gain a support of researchers (Danaher et al., 2017). From social scientists' vantage point, a holistic approach may be considered as the most fruitful.

1.2.9 Explainable artificial intelligence : a possible solution to algorithmic opacity

Although the notion of explainable artificial intelligence has been used in some circles for decades (Core et al., 2006), the contest recently launched by the Defense Advanced

[2] The inadvertent opacity may be considered as a kind of the contextual processes leading to the rise of cognitive overload. While Kirsh (2000) employed the individual- and team-level unit of analysis (e.g., a person and workplace), Stohl et al. (2016) adopted mostly a firm' level (i.e., a strategic perspective). In the present case, these processes are limited to the enterprises and other social structures with digital imprint, whereas the cognitive overload encompasses a broad family of conditions affecting human capacity to accomplish a task (e.g., Samson & Kostyszyn (2015) associate it with the decline of social trust).

Research Project Agency (DARPA, 2016) inspired researchers to take a closer look at the issue of inherent opacity in artificial intelligence. While the proliferation of diverse techniques and approaches would make it necessary to pursue a separate research project, analysis of a restricted number of recent papers pertaining to the explanation of AI can provide some ground for our discussion.

Table 1.6 Explainable artificial intelligence

Authors	Types & categories	Techniques
Biran & Cotton (2017)	Historical : expert systems and Bayesian networks; recommender systems; Current state: visualisation; prediction interpretation and justification; (inherently) interpretable models	Model-specific; model-free;
Doran et al., (2017)	Opaque: mechanism is not visible to user; Interpretable: human is able to explore and (mathematically) understand; Comprehensible: system provides human with the support to comprehend the path between various elements of system and results; Partially comprehensible: some events of such path are missing for human	Techniques for supporting external features such as confidence, trust, safety, fairness
Miller (2017)	Contrastive: it involves counterfactual cases; Selective: a limited range of cases for human to evaluate; Social: it involves human to human interaction and exchange of knowledge	Techniques for supporting conceptual frameworks, psychological processes, conversation
Launchbury (2017)	Contextual adaptation model involves: Perception; Learning (not in statistical terms); Abstraction; Reasoning while avoiding mistakes of the current wave (i.e. statistical learning)	

Table 1.6 presents different accounts of explainable artificial intelligence. Some observations from this table are worth highlighting. While Biran & Cotton (2017) and Doran et al. (2017) focus on the current and near-future state of affairs, Miller (2017) and Launchbury (2017) cover the long term. While explainable AI involves a large spectrum of techniques and approaches, explanation per se is a context-dependent social interaction. While the social nature of explanation necessitates the presence of individual actors and their groups, opaqueness and explanation are opposite concepts: the absence of one may mean the presence of the other. Moreover, opaqueness may be intermittent and may have gradients: appearing and disappearing if the same kind of explanation employs slightly different counterfactual stories presented to humans.

1.3 General Trends

1.3.1 Current trends in machine learning and artificial intelligence

New developments in machine learning such as capsule networks promise to improve deep learning models' performance. These improvements of building blocks might not substantially alter the relevance of learning algorithms, as the latter are extensions of the existing paradigm. Capsule networks will surely offer new possibilities in deep learning performance and push the boundaries of what is achievable in visual recognition tasks (Sabour et al., 2017). However, learning algorithms are still at their heart.

A generative model that is outside the deep learning mainstream is also worth discussing. Combining inspiration from neuroscience with different computer vision and probabilistic techniques, George et al. (2017) have succeeded in unifying vision and reasoning. Their efforts may be viewed as an extension of the connectionist

paradigm[3]. This combination of neural networks and probabilistic reasons may exhibit the same pattern of opacity that has haunted neural networks.

A more radical line in current research combines neural networks and artificial evolution : in short, a neuroevolution (Miikkulainen, 2015). Instead of looking for a feature among all the possible options, the goal is to identify a desirable behaviour without knowing in advance all the possibilities that might be present in the future. As in the previous example, learning algorithms play an important role because the combination of existing techniques and approaches depends on the learning algorithm and its possible mutations.

Finally, it is worth examining artificial intelligence and the pathways of its evolution. Experts disagree on the timeframe of AI's eventual arrival as well as on its future shape and behaviour (Sotala & Yampolskiy, 2017). This contingency is both worrisome and exciting for researchers and practitioners in economics, political sciences and business. The recent consensus on the safety of artificial general intelligence rests on the assumption that the pathway to artificial general intelligence will require a human-like cognitive architecture and embedded human values (Sotala & Yampolskiy, 2017, p. 71). The consensus also implies that social structures must be included in the definition of a learning algorithm. Indeed, these structures will ensure that evolving human values will be part of artificial general intelligence.

In this regard, the idea of hierarchical learning systems pursuing several goals may ensure that the transition from the narrow type of artificial intelligence presently in place to future general forms of artificial intelligence unfolds in an ethical manner. Formulated by Etzioni & Etzioni (2016; 2017) as "AI guardians," this approach falls under the more general category of "AGI nanny" in Sotala & Yampolskiy's typology

[3] The connectionist paradigm holds that the mammal brain' function and architecture are somehow captured by neural networks, an applied mathematics' object (Mira, 2008).

(2017, p. 64). All solutions to the ethical challenges of nascent AGI require that humans, as moral agents, be in the driving seat of the machine, which is itself an extension of human abilities. Therefore, human decision-making and the social context are necessary parts of the solution. The common theme to all of the above trends in deep learning and artificial intelligence research is the important role that human decision-making will need to play, even if assisted by learning algorithms.

1.3.2 Increasing autonomy of algorithmic systems

While deep learning used to be assisted by humans (i.e., labeled data, automatically coded rules, etc.), it is becoming increasingly independent (i.e., autonomous). For example, machine self-play and simulation-based techniques have reached a level of performance in the game of Go that human societies took thousands of years to achieve (Silver et al., 2016; Silver et al., 2017).

One possible implication is that deep learning will yield results that humans would take decades or centuries to obtain and to figure out what they are useful for. The same might be true of other unsupervised methods that could replace deep learning in the future. On the other hand, the energy (i.e., electric power) and other physical constraints required for such exploits may limit unsupervised methods' potential in real world settings for the coming years.

1.3.3 On the limits of understanding

In recent decades, expert voices have drawn attention to the difficulty of understanding the world, through scientific methods as much as through common sense. In 1975, Stent posed the question of the limits of scientific understanding of human nature, thereby challenging the dominant positivist doctrine of the time. Decades later, Barrow (1999) highlighted a recurrent pattern in various domains of research: a newly formulated explanation about a given phenomenon defines its own limits, thereby undermining public confidence about the progress of scientific endeavour. Other prominent

researchers have highlighted the limitations that come with reliance on any single type of knowledge, such as mathematics (Berreby, 2010). Facing such paradoxes, Arbesman (2016) proposes a middle ground between fear and awe and, as a practical step, advocates a return to generalist approaches to research. These approaches imply a pluralism of points of view combined with biological thinking (i.e., a mix of pragmatism and evolutionary ideas, following Charles Darwin).

It follows that the most rational and productive posture combines mild scepticism with a regular practice of questioning assumptions. It also demands that researchers embrace pragmatism and pluralism as signposts in the quest for understanding.

The overall message from this section is the following. The inherent opacity accompanying deep learning methods is a sufficiently novel phenomenon to warrant a call for a fresh perspective. Avoiding a purely technological point of view, this perspective should take into consideration broad trends in both machine learning and the social sciences.

CHAPTER II

OBJECTIVE AND APPROACH

2.1 Research questions

The main goal of this research project is to investigate the inherent opacity of learning algorithms and its eventual solutions by combining insights from several domains of research and fields of practice. The main question is the following: how can the opacity of learning algorithms be explained? In other words: (a) what are the events, actions, and activities as well as their progression in time behind algorithmic opacity? and (b) what are the most basic entities and their relationships that might be responsible for such opacity?

2.2 Methodology

The methodology for this report combines a qualitative literature review with concept development. This combination allowed us to identify and to adopt a new perspective (i.e., a predominantly sociological stance) while emphasizing an explanation-seeking nature of the whole project (Schryen et al., 2017). The qualitative literature review aims to identify, summarize and analyze available findings about the topic of interest by following the principles and guidelines proposed by Paré et al. (2015) and Templier &

Paré (2015). These are: selective search strategy, broad scope, multiple sources and a narrative chronological synthesis. Categories of search terms consisted of algorithm, opacity, literature reviews, deep learning, artificial intelligence and their variations. The review was conducted in multiple iterative phases during June/December 2017.

Peer-reviewed journal articles and conference papers were sourced from the scholarly databases' providers, namely SCOPUS, Web of Science, ACM, EBSCO, and ProQuest. Additionally, arXiv, bioArxiv, SSRN and other publicly accessible sites were sources of supplemental documents, qualified as "gray literature." Moreover, the "fit for purpose" criteria suggested by Adams et al. (2017) were used to evaluate the quality of documents before subjecting them to further analysis. These criteria are: i) author previously published in peer-reviewed outlets and ii) authors' research institution affiliation. The CLUSTER technique proposed by Booth et al. (2013) was used to identify additional discourses, themes, and streams of research in order to enrich the overall collection of concepts, constructs, theories, and metaphors. CLUSTER is an acronym for "citations, leading authors, unpublished works, theories, early examples, and related projects." This manually executed technique enlarges the scope of sources and documents for review. The most prominent citations in the source documents were located to identify publications, authors, projects, programs and institutions in an iterative manner. Their disparate natures (i.e., sources and documents) come at the price of formal quality assessment, which is almost impossible to conduct.

The concept development phase was conducted using a modified hybrid approach, as defined by Branch & Rocchi (2015, pp. 128-129). The word "hybrid" refers to the combination of: i) qualitative analysis of the overall collection (used to avoid possible contradictions), ii) refinement of relevant core elements of plausible explanatory schemas, and iii) their combination in order to present a consistent and simple explanation. The coherence of the final account and the plausibility of its constitutive components characterize the guiding principles of this phase. Finally, investigating the

opacity of learning algorithms requires us to identify the most basic elements and dynamics giving rise to such opacity. The combination of qualitative methods (i.e., a narrative form of literature review and concept development) allowed us to obtain the following results.

CHAPTER III

DEVELOPING AN EXPLANATION

This section begins with an analysis of the origins of algorithmic opacity, it continues with a presentation of its mechanisms and concludes with a short history of algorithmic opacity.

3.1 The origins of algorithmic opacity

In light of Hill's (2016) definition of an algorithm, three possible sources of algorithmic opacity may be advanced: descriptive, contextual and purposeful.

Descriptive opacity may be formulated as the difference between a machine's execution of an algorithm and the human understanding of such a process, as pertains to its description. The same idea makes it possible to identify contextual and purposeful opacity, as applied to the algorithm's context and purpose, respectively.

For illustrative purposes, let us consider a case of descriptive opacity: let us imagine a situation where social actors in an enterprise engage in activities leading to the creation and use of an algorithm. For example, a team's goal is to construct and deploy an algorithmic system by identifying what this system will do and by describing its behavior in qualitative and quantitative terms. In this process, various actors within the enterprise will contribute to the definition of an algorithm and their choice will be embedded in the artefacts constituting the description. From a management and

organizational perspective, this decision-making process requires the actors to deal with conflict, tensions, divergences of opinions and other impediments to consensus about the elements that will constitute the description of the future algorithmic system. From a technological point of view, this description will include professionally bounded artefacts (e.g., various models and structures, visual representations, pseudo-codes and human-produced codes). Based on the knowledge and practice that ICT workers bring to the table, the mutual efforts will lead to the execution and deployment of the algorithmic system. Analogous illustrative accounts may be developed for the two other elements of the definition of an algorithm, namely, for the algorithm's context and purpose. Although similar, the account of algorithms' purpose includes mostly human-specific attributes, such as intentionality.

Finally, the numerous sub-cases of overlap between description, purpose and context may considerably increase the varieties of opacity. Given the number of sub-cases to explore and elements to isolate, and adding the temporal dimension to the picture, it will be easier to investigate opacity by considering it as a process of interaction between all these different elements and components.

3.2 The comparison of proposed and existing typologies

It is worth comparing current typologies of opacity and the one proposed in this section. According to Burrell (2016, pp. 3-4), intentional opacity arises when an institutional actor wilfully hides information in protection of its interest. By contrast, the purposeful opacity here proposed emerges naturally, as people engage with algorithmic systems. In such a case, even if the corporate decision is to obfuscate, the intention and action of individuals may override such corporate aims. Users, clients, partners and regulators, whose actions and intention are clearly outside of the reach of a particular firm, may

nullify the corporate aim as their actions and decision making impact the learning algorithm.

Additionally, there is a slight difference in the meaning between purposeful and intentional opacity as described in this paper and in the work of Burrell (2016) respectively. Although rhetorically they are very close, purposeful opacity includes intentional opacity, but not vice-versa. In this work, purposeful opacity incorporates the purposes that ordinary users of algorithm expect from the algorithm. Casual, regular or unsuspected users presume a certain behaviour of algorithmic systems that is in-line with their own desires and objectives. In Burrell's work, the only intentions accounted for are those of corporate stakeholders (i.e. creators, operators and owners of algorithmic systems).

Moreover, contextual opacity and its role are absent from the work of Burrell (2016). In the present text, contextual opacity is what is not explicitely included in purposeful and descriptive opacity. For example, the data that feed into algorithmic systems during their operations and the various human choices and actions captured by such data belong to this category. In the case of traditional algorithms, the absense of this category did not pose a problem, as humans remained in charge of modifying the structure of traditional algorithms, and as the boundaries between algorithm and data were static and clear. In a regime of learning algorithms, these constraints do not hold anymore. This does play some role in the rise and persistence of opacity. Even if descriptive and purposeful opacity were somehow resolved, contextual opacity would still persist. And given the distinct characteristics of learning algorithms (see chapter II), descriptive and purposeful opacity may reappear.

Finally, as proposed here, descriptive opacity takes into account two well-established types of opacity, namely opacity as "technical illiteracy" and opacity as mismatch between mathematical formula and the human interpretation of such formula (Burrell, 2016, pp. 4-5). Descriptive opacity encompasses both these types as they are just

different stages and facets in the process of creating and using algorithms from the expert's point of view. Moreover, descriptive opacity includes also what regular users, the ones who are not professionals trained in the art of machine learning, may see, understand, expect and do with the algorithmic system. Their behaviour will be captured by algorithms through 'digital crumbs', pieces of information attached to actions such as sending an email or buying a cup of coffee (Pentland, 2013).

3.3 The mechanism of algorithmic opacity

The overall mechanism of algorithmic opacity involves several basic elements, constraints and conditions. It possesses two basic dimensions: the social (i.e., entities, their properties and their actions (Elster, 2015)) and the algorithmic (i.e., the dimension that is of interest to ICT researchers investigating the inner workings of an algorithm). Furthermore, as the object of supplementary refinement and in order to be qualified as such, the mechanism requires the presence of the following elements:

The first, called "social entity", is made up of human beings arranged into groups. Their actions, beliefs and intentions define the description and the context of an algorithmic system. The second element of the mechanism corresponds to the humans' properties: i.e., their aims, intentions and purposes. These human properties influence the description, purpose and context of an algorithmic system. The last element corresponds to the actions that individuals and their groups perform in order to achieve their goals, aspirations and intentions. These actions determine the context, purpose and description of an algorithmic system.

Another element of these mechanisms requires human decision-making in regard to the situations (i.e., contextual components) where and when it could be realised. This

decision-making process involves various items of an algorithmic system tied to the social entities, their properties and their actions. For example, the various symbols of mathematical formulae and the choices of models of data to be fed into algorithmic systems all belong to the description and the context of a learning algorithm. Competing interpretations of models, entities and data may serve as indicators of rising algorithmic opacity. This opacity would otherwise not be directly observable to the actors involved in the process.

The third element involves interpretation of the results of the algorithm executed as well as human understanding of these results. Intrinsically personal, often intangible and fallible by nature, this interpretation appears to be highly subjective insofar as it is perceived by external actors and its activities are remarkably context-dependent.

Finally, the overall mechanism of algorithmic opacity is context-dependent: the individual elements and components of the mechanism, as well as their interplay, vary from situation to situation. However, the overall process and logic remain the same.

3.4 The history of algorithmic opacity

The history of algorithmic opacity is the story of algorithmic systems' interaction with various groups of people during their life time. It is a story of gradual and subtle accumulation of small, invisible units of opaqueness, defined by human decisions (i.e., design, development and operation). Each member of a group, with their beliefs and assumptions about the world around them, makes decisions that somehow contribute to the accumulation of opacity. This accumulation of opacity is not unlimited and it emerges once a threshold is attained.

In the case of perceptions (e.g., image recognition and voice generation), this threshold is easily identifiable by humans. For example, it is difficult to admit as real a picture of an elephant identified by deep learning algorithm as a household cat even if accompanied with a high level of statistical confidence (McDaniel et al., 2016). This would not be the case in situations where historical and evolutionary markers were not present, nor in less familiar situations or in those requiring longstanding experience or dexterity. For example, molecules and atoms do not possess such easily detectable indicators as the evolution of humans as a biological species and the history of human society circumscribe and sharpen for millions and thousands of years (Gawehn et al., 2016). Thus, life-science researchers and practitioners resort to the combination of deep learning with other techniques and approaches in order to delineate the use of deep machine learning models (Hoehndorf et al., 2017; Jiang et al., 2017).

This overall mechanism could also explain the elusive nature of algorithmic opacity (i.e., not only in the case of deep learning according to Ojha et al. (2017)). Once isolated from their environmental contexts, the individual components or elements of the mechanism might not be reproducible elsewhere or even identified as such. And even if accidentally identified and subsequently associated with numbers or other concepts, they would not be directly observable in different situations or various contexts. Moreover, the fluid nature of algorithmic opacity in the case of deep learning might be observed in other machine learning techniques or in up-to-date variations of Artificial Intelligence once this one loses its economic lustre or falls out of fashion in the eyes of society.

If one accepts the existence of such a mechanism, the following situation arises. On the one hand, the most problematic characteristic of a learning algorithm is its irreversibility. Once the algorithm is executed and once the results are in front of a human, it is extremely difficult, if not impossible, to go back in time with the goal of reconstructing the whole system (i.e., the description, purpose and context). While with

traditional algorithms, some techniques, such as reverse-engineering (Canfora et al., 2011), allow researchers to recreate the description and, with additional effort, the whole system, learning algorithms do not permit such luxury because of continuous interactions with human beings. To reproduce an initial (i.e., before execution or deployment) learning algorithm, one would need to reconstruct all the elements and components that constitute its description, context and purpose, all of which are influenced by previous human decision-making. Given the current scale and pervasiveness of these systems, and given the fluid and changing nature of the context and description, these efforts would require the alignment of a rare set of circumstances not happening spontaneously. For example, serving as a mediating passage point between the social and machine realm, Big Data belongs to a context (Loebbecke & Picot, 2015) while the deep learning, to description. Together, they constitute an amalgam guaranteeing such irreversibility.

On the other hand, the presence of such a mechanism ensures that, even if the learning algorithm could be temporarily stopped and reversed, continuous human decision-making could invalidate such efforts. These efforts can quickly seem futile if one does not acknowledge that avoiding opacity requires implementing something else, namely its opposite, transparency or its alternative, human decision making. Aside from the technological means of building transparency and once humans are accounted for, the best way to avoid algorithmic opacity, one might presume, is to ensure that human decision-making assisted by a learning algorithm is consistently improved. Moreover, given the paradox of transparency (section 1.2.8, pp. 30-31) which states that the institutionalized efforts for increasing transparency could lead to greater opacity, it would be natural to act on the basis of both of these situations. The common jonction point in the chain of actions leading to the opacity of an algorithmic system and the paradox of transparency is the human decision making in an enterprise setting. If human decision-making is the main culprit behind the rise of algorithmic opacity, then it is necessary to create and to maintain conditions in which decision-making does not

lead to greater opacity. Thus, if humans' decisions relied heavily on learning algorithms, one would need to make sure that the algorithmic systems in question were transparent (i.e., understandable to humans) in the technological sense. The paradox of opacity, transparency, human decision-making and learning algorithms is that, while it is impossible to go back in time with learning algorithms, one should constantly exercise and maintain transparency and the quality of human decision-making. In turn, the quality of that decision-making depends on the degree of opacity or transparency of the learning algorithm.

CHAPTER IV

CONCLUSION

4.1 Implications

The most obvious implication of learning algorithms' opacity for social scientists is the following. The proposed model of algorithmic definition may be useful for clearly distinguishing between algorithms and other digital artifacts because it explicitly considers the social dimension of algorithms. Specifically, because the definition describes algorithms in relation to human objectives and purposes, it brings focus to beliefs, intentions and judgments as human-specific characteristics. Moreover, various forms of algorithms' description admit the ones easily understandable by most people instead of confining it to a selected few. Finally, the definition's emphasis on context recognizes the influence of an algorithm's environment, which includes human beings and their interactions.

These results may enhance previous investigations into algorithmic systems. For example, Kitchin (2016) proposed a range of techniques, which gained the endorsement of the research community (Danaher et al., 2017). Once elaborated, our model of learning algorithm opacity could provide validation methods for the proposed techniques.

The definitions of algorithms and learning algorithms presented in this paper, together with the process-oriented conceptions of opacity, may be of interest to researchers in

various fields. They provide analytical methods for enhancing the transparency of the decision-making processes involved in the building and use of algorithmic systems. In light of the attention that the public, policymakers and other social actors have been giving to digital platforms in the social media and communication landscape, it seems reasonable to suppose that these actors will demand a justification of the appropriateness and adequacy of efforts aiming to avoid the opacity of newly built or existing digital platforms. As far as other features of algorithmic systems, such as autonomy, are concerned, the definitions of algorithms and learning algorithms given in this paper may serve as the bases for fruitful discussion among researchers from legal, medical and other backgrounds.

4.2 Limitations

Possible objections to the findings by researchers and practitioners in ICT might be countered from various perspectives. The perspectives of social scientists and MOS scholars concern mostly the results, inputs and environments of algorithmic systems. The questions they ask are of two types: first, about the effects of algorithms; second, about ways to affect algorithmic systems. From their technological perspective, ICT researchers and practitioners are preoccupied with the internal components of algorithms, their structures and their dynamics. In other words, both groups consider the same object, but from different positions. Moreover, an axiomatic type of definition – like the one pursued by computer scientists – does not fundamentally contradict the one proposed in this work. Another way to look at these two perspectives is to consider them as complementary. In view of avoiding controversies and misunderstandings, collaboration between researchers may be the best way to go.

As a standalone approach following a qualitative tradition in MOS, concept development might have followed the guidelines of Podsakoff et al. (2016). However, given the emergent nature of learning algorithms, the disparate voices on the topic of algorithmic opacity and the difficulty defining learning algorithms in any non-technological sense, it may have been counterproductive to assiduously follow these guidelines. Moreover, some of the elements of these guidelines are implicitly present in this work. For example, the use of a literature review, the identification of core attributes and the initial description of concepts' properties correspond to this work artefacts and phases.

The extensive use of gray literature is justified by the triangulation principle at multiple levels. The diversity of sources comes at a price, however: namely, the difficulty of drawing an overall coherent picture about learning algorithms' opacity. Moreover, this diversity also precludes the application of quality assessment methods such as those suggested by Podsakoff et al. (idem.) and Paré et al. (2015).

This report's use of techniques and approaches from related domains may raise some questions about the trustworthiness of its results. The qualification label of narrative type of review may be justified by the fact that MOS and IS share a common conceptual background, as Baskerville & Myers (2002) and Iivari (2017) have affirmed. Moreover, by bibliometrical measures, both domains are very closely related in the universe of social sciences and technology (Börner et al., 2012; Leydesdorff et al., 2013).

4.3 Future work

Future research will aim to refine current understandings of learning algorithms and the overall mechanism responsible for algorithmic opacity. It will divide into several independent streams.

Examples of the first stream of research approaches include factor analysis (i.e., exploratory or confirmatory). These approaches will identify individual factors, variables or constituent elements of the mechanism. The combination of grounded-theory methods, factor analysis and case studies will necessitate collaboration with business partners in an industrial setting.

The second stream offers a panoply of options some of which are as follows. The focus on clarifying learning algorithms by following the guidelines given by Podsakoff et al. (2016) may also include simulation of agent-based models, as suggested by Bruch & Atwell (2015). This simulation may uncover possible explanation schemas not linked with empirical facts (Hedström & Ylikoski, 2010, pp. 62-63). Moreover, these efforts will also include the clarification of the distinction between the opacity of learning algorithms and other closely related attributes, such as transparency, accountability, fairness, and interpretability. This list may eventually comprise other qualitative characteristics all of which are necessary for providing a rich theoretical foundation aiming at explanation and understanding of algorithmic systems' behavior.

The goal of refining the conceptual basis is to identify in specific terms what groups of people or individuals participate in the mechanism. It also aims to address related questions pertaining to the actions and interactions that contribute to the emergence of algorithmic opacity in particular contexts. More specifically, according to Hedström & Ylikoski (idem., pp. 59), the general mechanism may be situational or transformational in form.

In more practical terms, this second stream of research may take the form of multiple case studies in enterprises or other social organizations. These case studies may include a combination of semi-structured interviews and focus groups as primary data-gathering methods for the evaluation and validation of the mechanism's effects. Several groups of experts could be involved in this evaluation. More specifically, the Delphi method might be a good starting point of such investigation, as the communities

of researchers specialized in deep learning and members of 'explainable AI' movement are highly concentrated and relatively easy to identify. Furthermore, the multiple-case studies may need to include a broad range of industrial sectors as well as geographical locations.

For computer scientists and statisticians, the interpretation of mathematical symbols and formulae will make it possible to highlight the difference between what machine executes and what human understand and reveal compromises and trade-offs in cases where humans are dealing with the theoretical structures of learning algorithms' mathematical aspects. When learning algorithms are prepared for real-world operation, interviews with software engineers and ICT people will make it possible to discover controversies and disagreements. The focus will be on the accumulation of elements from previous algorithm' development steps that play some role in the emergence of algorithmic opacity, in addition to this step's trade-offs, paradoxes and dilemma.

As for business-minded people and their customers, real-world applications of learning algorithms could potentially reveal the same patterns of mismatches in interpretations and unmask paradoxes and compromises. In all these settings, human decision-making (as a small-group practice) could be measured with proper procedures: for example, formative or reflective measurement models (Johnson et al., 2011). In fact, a reflective measurement model would arguably be most appropriate, as the processes and activities responsible for the rise of opacity could be assessed empirically via their effects (Rigdon, 2016; Sarstedt et al., 2016). Moreover, the collected data may also be used to discover and evaluate individual elements of the explanatory mechanism, such as criteria, conditions and aftermath, thus helping to alleviate the effects of algorithmic opacity.

The outcomes of these multiple case studies will include best practice suggestions, guidelines, and indicators for practitioners and decision-makers. These outcomes will enable the stakeholders to detect and manage possible emergence of algorithmic

opacity. Moreover, these practice-oriented outcomes will include a description of situations and settings conducive to the emergence of algorithmic opacity as well as recommendations on how to handle those cases.

REFERENCES

Adams, R. J., Smart, P. & Huff, A. S. (2017). Shades of Grey: Guidelines for Working with the Grey Literature in Systematic Reviews for Management and Organizational Studies. *International Journal of Management Reviews*, *19*(4), 432-454. doi: 10.1111/ijmr.12102

Ambrose, M. L. (2015). Lessons from the avalanche of numbers: big data in historical perspective. *I/S: A Journal of Law and Policy for the Information Society*, *11*(2), 201-279.

Ananny, M. & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *0*(0), 146144481667664. doi: 10.1177/1461444816676645

Arbesman, S. (2016). *Overcomplicated: Technology at the Limits of Comprehension.* New York, NY, USA : Current.

Armando, V. (2017). Business Applications of Deep Learning. In K. Pradeep et T. Arvind (eds.), *Ubiquitous Machine Learning and Its Applications* (p. 39-67). Hershey, PA, USA : IGI Global.

Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. (2017). Deep Reinforcement Learning: A Brief Survey. *IEEE Signal Processing Magazine*, *34*(6), 26-38. doi: 10.1109/MSP.2017.2743240

Barreto, J. M. & de Azevedo, F. M. (1993). Connectionist expert systems as medical decision aid. *Artificial Intelligence in Medicine*, *5*(6), 515-523. doi: 10.1016/0933-3657(93)90041-Z

Barrow, J. D. (1999). *Impossibility: The Limits of Science and the Science of Limits.* New York, NY, USA : Oxford University Press. Retrieved from https://books.google.ca/books?id=2J3mCwAAQBAJ

Baskerville, R. L. & Myers, M. D. (2002). Information Systems as a Reference Discipline. *MIS Quarterly*, *26*(1), 1-14. doi: 10.2307/4132338

Bechmann, A. (2017). *Keeping it Real: From Faces and Features to Social Values in Deep Learning Algorithms on Social Media Images : Proceedings of the 50th Hawaii International Conference on System Sciences January 4 - 7, 2017.* Honolulu, HI : University of Hawai'i at Manoa. doi: 10.24251/HICSS.2017.218

Beer, D. (2016). The social power of algorithms. *Information, Communication & Society*, *20*(1), 1-13. doi: 10.1080/1369118X.2016.1216147

Bengio, Y., Simard, P. & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks, 5*(2), 157-166. doi: 10.1109/72.279181

Berreby, D. (2010). The Limits of Understanding. *The Scientist* (Midland, ON, Canada), p. 8-10.

Biegel, B. & Kurose, J. F. (2016). *The National Artificial Intelligence Research and Development Strategic Plan.* Washington, DC, USA : White House Office of Science and Technology Policy (OSTP). Retrieved from https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/national_ai_rd_strategic_plan.pdf

Binns, R. (2017). Algorithmic Accountability and Public Reason. *Philosophy & Technology*, 1-14. doi: 10.1007/s13347-017-0263-5

Biran, O. & Cotton, C. (2017). *Explanation and justification in machine learning: A survey : Proceeding of IJCAI-17 Workshop on Explainable AI (XAI), 20 August 2017.* Los Angeles, CA : International Joint Conferences on Artificial Intelligence Corp. Retrieved from http://www.intelligentrobots.org/files/IJCAI2017/IJCAI-17_XAI_WS_Proceedings.pdf#page=8

Bolin, G. & Schwarz, J. A. (2015). Heuristics of the algorithm: Big Data, user interpretation and institutional translation. *Big Data & Society*, *2*(2). doi: 10.1177/2053951715608406

Booth, A., Harris, J., Croot, E., Springett, J., Campbell, F. & Wilkins, E. (2013). Towards a methodology for cluster searching to provide conceptual and contextual "richness" for systematic reviews of complex interventions: case study (CLUSTER). *BMC Medical Research Methodology, 13*(1), 118. doi: 10.1186/1471-2288-13-118

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., Larivière, V, & Boyack, K. W. (2012). Design and Update of a Classification System: The UCSD Map of Science. *PLOS ONE, 7*(7), e39464. doi: 10.1371/journal.pone.0039464

Branch, J. & Rocchi, F. (2015). Concept Development: A Primer. *Philosophy of Management, 14*(2), 111-133. doi: 10.1007/s40926-015-0011-9

Bruch, E. & Atwell, J. (2015). Agent-Based Models in Empirical Social Research. *Sociological Methods & Research, 44*(2), 186-221. doi: 10.1177/0049124113506405

Bucher, T. (2017). The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society, 20*(1), 30-44. doi: 10.1080/1369118X.2016.1154086

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1-12. doi: 10.1177/2053951715622512

Canfora, G., Penta, M. D. & Cerulo, L. (2011). Achievements and challenges in software reverse engineering. *Communications of the ACM, 54*(4), 142-151. doi: 10.1145/1924421.1924451

Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S. & Rao, R. M. (2017). Interpretability of Deep Learning Models: A Survey of Results. In A. Swami, C. Williams, D. Verma, G. Pearson & T. Pham (eds.), *IEEE Smart World Congress 2017 Workshop: DAIS 2017 - Workshop on Distributed Analytics InfraStructure and Algorithms for Multi-Organization Federations, 7-8 August 2017.* Wales, UK : Cardiff University. Retrieved from http://orca.cf.ac.uk/id/eprint/101500

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., . . . Greene, C. S. (2017). *Opportunities And Obstacles For Deep Learning In Biology And Medicine.* (version 1). Retrieved from https://dx.doi.org/10.1101/142760

Core, M. G., Lane, H. C., Van Lent, M., Gomboc, D., Solomon, S. & Rosenberg, M. (2006). *Building explainable artificial intelligence systems : Eighteenth Conference on Innovative Applications of Artificial Intelligence (IAAI-06), July 16–20, 2006.* Palo Alto, CA : The AAAI Press. Retrieved from https://aaai.org/Papers/AAAI/2006/AAAI06-293.pdf

DARPA. (2016). *Explainable Artificial Intelligence (XAI).* DARPA-BAA-16-53. Arlington, VA, USA : Defense Advanced Research Projects Agency, Government of US, USA. Retrieved from https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, *29*(3), 245-268. doi: 10.1007/s13347-015-0211-1

Danaher, J., Hogan, M. J., Noone, C., Kennedy, R., Behan, A., Paor, A. D., . . . Shankar, K. (2017). Algorithmic governance: Developing a research agenda through the power of collective intelligence. *Big Data & Society*, *4*(2), 2053951717726554. doi: 10.1177/2053951717726554

Datta, A., Sen, S. & Zick, Y. (2016). *Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems : 2016 IEEE Symposium on Security and Privacy (SP), 22-26 May 2016.* Washington, DC ; IEEE Computer Society. doi: 10.1109/SP.2016.42

Dhar, V. (2016). The Future of Artificial Intelligence. *Big Data*, *4*(1), 5-9. doi: 10.1089/big.2016.29004.vda

Diakopoulos, N. (2014). *Algorithmic accountability reporting: On the investigation of black boxes.* Tow Center for Digital Journalism, Columbia University. New York, NY, USA : Columbia University. Retrieved from https://doi.org/10.7916/D8ZK5TW2

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59*(2), 56-62. doi: 10.1145/2844110

Diakopoulos, N. (2017). Enabling Accountability of Algorithmic Media: Transparency as a Constructive and Critical Lens. In T. Cerquitelli, D. Quercia & F. Pasquale (eds.), *Transparent Data Mining for Big and Small Data* (pp. 25-43). Cham, Switzerland : Springer International Publishing.

Doran, D., Schulz, S. & Besold, T. R. (2017). *What Does Explainable AI Really Mean? A New Conceptualization of Perspectives.* (version 1). Retrieved from https://arxiv.org/abs/1710.00794v1

Ekbia, H., Mattioli, M., Kouper, I., Arave, G., Ghazinejad, A., Bowman, T., . . . Sugimoto, C. R. (2015). Big data, bigger dilemmas: A critical review. *Journal of the Association for Information Science and Technology, 66*(8), 1523-1545. doi: 10.1002/asi.23294

Eslami, M., Karahalios, K., Sandvigt, C., Vaccaro, K., Rickman, A., Hamilton, K. & Kirlik, A. (2016). *First I "like" it, then I hide it: Folk Theories of Social Feeds : Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* New York, NY : ACM, Inc. doi: 10.1145/2858036.2858494

Etzioni, A. & Etzioni, O. (2016). Designing AI systems that obey our laws and values. *Communications of the ACM, 59*(9), 29-31. doi: 10.1145/2955091

Etzioni, A. & Etzioni, O. (2017). Incorporating Ethics into Artificial Intelligence. *The Journal of Ethics, 21*(4), 403-418. doi: 10.1007/s10892-017-9252-2

Floridi, L. (2009). The Information Society and Its Philosophy: Introduction to the Special Issue on "The Philosophy of Information, Its Nature, and Future Developments". *The Information Society, 25*(3), 153-158. doi: 10.1080/01972240902848583

Floridi, L. (2014). *The Fourth Revolution: How the Infosphere is Reshaping Human Reality.* Oxford, UK : Oxford University Press. Retrieved from https://books.google.ca/books?id=ruL2AwAAQBAJ

Gallant, S. I. (1988). Connectionist expert systems. *Communications of the ACM*, *31*(2), 152-169. doi: 10.1145/42372.42377

Gawehn, E., Hiss, J. A. & Schneider, G. (2016). Deep Learning in Drug Discovery. *Molecular Informatics*, *35*(1), 3-14. doi: 10.1002/minf.201501008

Geiger, R. S. (2017). Beyond opening up the black box: Investigating the role of algorithmic systems in Wikipedian organizational culture. *Big Data & Society*, *4*(2), 1-14. doi: 10.1177/2053951717730735

George, D., Lehrach, W., Kansky, K., Lázaro-Gredilla, M., Laan, C., Marthi, B., . . . Phoenix, D. S. (2017). A generative vision model that trains with high data efficiency and breaks text-based CAPTCHAs. *Science*, *358*(6368), eaag2612. doi: 10.1126/science.aag2612

Gillespie, T. (2017). Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information, Communication & Society*, *20*(1), 63-80. doi: 10.1080/1369118X.2016.1199721

Goertzel, B. (2014). Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence, 5*(1), 1-48. doi: 10.2478/jagi-2014-0001

Goh, G. B., Hodas, N. O. & Vishnu, A. (2017). Deep learning for computational chemistry. *Journal of Computational Chemistry*, *38*(16), 1291-1307. doi: 10.1002/jcc.24764

Goodfellow, I., Bengio, Y. & Courville, A. (2016). *Deep learning*. Cambridge, MA : MIT Press. Retrieved from https://books.google.ca/books?id=Np9SDQAAQBAJ

Gurevich, Y. (2012). What Is an Algorithm? In M. Bieliková, G. Friedrich, G. Gottlob, S. Katzenbeisser et G. Turán (eds.), *SOFSEM 2012: Theory and Practice of Computer Science: 38th Conference on Current Trends in Theory and Practice of Computer Science, Špindlerův Mlýn, Czech Republic, January 21-27, 2012.*

Proceedings (pp. 31-42). Berlin, Heidelberg : Springer Berlin Heidelberg. doi: 10.1007/978-3-642-27660-6_3

Gurevich, Y. (2015). Semantics-to-Syntax Analyses of Algorithms. In G. Sommaruga & T. Strahm (eds.), *Turing's Revolution: The Impact of His Ideas about Computability* (pp. 187-206). Cham, Switzerland : Springer International Publishing. doi: 10.1007/978-3-319-22156-4_7

Hedström, P. & Ylikoski, P. (2010). Causal Mechanisms in the Social Sciences. *Annual Review of Sociology, 36*(1), 49-67. doi: 10.1146/annurev.soc.012809.102632

Hill, R. K. (2016). What an Algorithm Is. *Philosophy & Technology, 29*(1), 35-59. doi: 10.1007/s13347-014-0184-5

Hinton, G. E., Osindero, S. & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation, 18*(7), 1527-1554. doi: 10.1162/neco.2006.18.7.1527

Hoare, C. A. R. (1969). An axiomatic basis for computer programming. *Communications of the ACM, 12*(10), 576-580. doi: 10.1145/363235.363259

Hoehndorf, R., Queralt-Rosinach, N. & Kuhn, T. (2017). Data Science and symbolic AI: Synergies, challenges and opportunities. *Data Science, 1*(1-2), 1-12. doi: 10.3233/ds-170004

Iivari, J. (2017). Information system artefact or information system application: that is the question. *Information Systems Journal, 27*(6), 753-774. doi: 10.1111/isj.12121

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., . . . Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology, 2*(4), 230-243. doi: 10.1136/svn-2017-000101

Johnson, R. E., Rosen, C. C. & Chang, C.-H. (2011). To Aggregate or Not to Aggregate: Steps for Developing and Validating Higher-Order Multidimensional Constructs. *Journal of Business and Psychology*, *26*(3), 241-248. doi: 10.1007/s10869-011-9238-1

Jordan, M. I. & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. doi: 10.1126/science.aaa8415

Kerr, N. L. & Tindale, R. S. (2004). Group Performance and Decision Making. *Annual Review of Psychology,* *55*(1), 623-655. doi: 10.1146/annurev.psych.55.090902.142009

Kitchin, R. (2016). Thinking critically about and researching algorithms. *Information, Communication & Society*, *20*(1), 14-29. doi: 10.1080/1369118X.2016.1154087

Kirsh, D. (2000). A Few Thoughts on Cognitive Overload. *Intellectica, 1*(30), 19-51.

Kowalski, R. (1979). Algorithm = logic + control. *Communications of the ACM*, *22*(7), 424-436. doi: 10.1145/359131.359136

L'Heureux, A., Grolinger, K., Elyamany, H. F. & Capretz, M. A. M. (2017). Machine Learning With Big Data: Challenges and Approaches. *IEEE Access*, *5*, 7776-7797. doi: 10.1109/ACCESS.2017.2696365

Launchbury, J. (2017). *A DARPA Perspective on Artificial Intelligence*. Retrieved December 19, 2017 from https://www.darpa.mil/about-us/darpa-perspective-on-ai

LeCun, Y., Bengio, Y. & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444. doi: 10.1038/nature14539

Lepri, B., Oliver, N., Letouzé, E., Pentland, A. & Vinck, P. (2017). Fair, Transparent, and Accountable Algorithmic Decision-making Processes. *Philosophy & Technology*. doi: 10.1007/s13347-017-0279-x

Lester, M. (2018). The Creation and Disruption of Innovation? Key Developments in Innovation as Concept, Theory, Research and Practice. In T. Clarke et K. Lee (eds.), *Innovation in the Asia Pacific: From Manufacturing to the Knowledge Economy* (pp. 271-328). Singapore : Springer Singapore.

Leydesdorff, L., Carley, S. & Rafols, I. (2013). Global maps of science based on the new Web-of-Science categories. *Scientometrics*, *94*(2), 589-593. doi: 10.1007/s11192-012-0784-8

Lin, H. W., Tegmark, M. & Rolnick, D. (2017). Why Does Deep and Cheap Learning Work So Well? *Journal of Statistical Physics*, *168*(6), 1223-1247. doi: 10.1007/s10955-017-1836-5

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y. et Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing, 234*, 11-26. doi: 10.1016/j.neucom.2016.12.038

Loebbecke, C. & Picot, A. (2015). Reflections on societal and business model transformation arising from digitization and big data analytics: A research agenda. *The Journal of Strategic Information Systems, 24*(3), 149-157. doi: 10.1016/j.jsis.2015.08.002

McCarthy, J., Minsky, M. L., Rochester, N. & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence, August 31, 1955. *AI magazine*, *27*(4), 12-14. doi: 10.1609/aimag.v27i4.1904

McDaniel, P., Papernot, N. & Celik, Z. B. (2016). Machine Learning in Adversarial Settings. *IEEE Security & Privacy, 14*(3), 68-72. doi: 10.1109/MSP.2016.51

Miikkulainen, R. (2015). Evolving Neural Networks. In Jiménez-Laredo, J. (ed.), *Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation, July 11 - 15, 2015* (pp. 137-161). 2756577 New York, NY : ACM. doi: 10.1145/2739482.2756577

Miller, T. (2017). *Explanation in artificial intelligence: Insights from the social sciences.* (version 1). Retrieved from https://arxiv.org/abs/1706.07269v1

Miotto, R., Wang, F., Wang, S., Jiang, X. & Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, 1-11. doi: 10.1093/bib/bbx044

Mira, J. M. (2008). Symbols versus connections: 50 years of artificial intelligence. *Neurocomputing, 71*(4), 671-680. doi: 10.1016/j.neucom.2007.06.009

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 2053951716679679. doi: 10.1177/2053951716679679

Moravčík, M., Schmid, M., Burch, N., Lisý, V., Morrill, D., Bard, N., . . . Bowling, M. (2017). DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science, 356*(6337), 508-513. doi: 10.1126/science.aam6960

Moschovakis, Y. N. (1998). On founding the theory of algorithms. In H. G. Dales & G. Oliveri (eds.), *Truth in mathematics* (pp. 71-104). New York, NY : Clarendon Press.

Moschovakis, Y. N. (2001). What Is an Algorithm? In B. Engquist & W. Schmid (eds.), *Mathematics Unlimited — 2001 and Beyond* (pp. 919-936). Berlin, Heidelberg : Springer Berlin Heidelberg. doi : 10.1007/978-3-642-56478-9_46

Müller, V. C. & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In C. V. Müller (ed.), *Fundamental Issues of Artificial Intelligence* (pp. 553-570). Cham, Switzerland : Springer International Publishing.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R. & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1-21. doi: 10.1186/s40537-014-0007-7

National Research Council, (2013). *Frontiers in Massive Data Analysis*. Washington, DC, USA : The National Academies Press. doi: 10.17226/18374

Ojha, V. K., Abraham, A. & Snášel, V. (2017). Metaheuristic design of feedforward neural networks: A review of two decades of research. *Engineering Applications of Artificial Intelligence, 60*, 97-116. doi: 10.1016/j.engappai.2017.01.013

Olden, J. D. & Jackson, D. A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, *154*(1), 135-150. doi: 10.1016/S0304-3800(02)00064-9

Olleros, F. J. & Zhegu, M. (2016a). *Research Handbook on Digital Transformations*. Cheltenham, Gloucestershire, UK : Edward Elgar Publishing.

Olleros, F. J. & Zhegu, M. (2016b). Digital Transformations: An Introduction. In F. J. Olleros & M. Zhegu (eds.), *Handbook of Research on Digital Transformations* (pp. 1-19). Cheltenham, Gloucestershire, UK : Edward Elgar Publishing.

Paré, G., Trudel, M.-C., Jaana, M. & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, *52*(2), 183-199. doi: 10.1016/j.im.2014.08.008

Pentland, A. (2013). The data-driven society. *Scientific American, 309*(4), 78-83

Perel, M. & Elkin-Koren, N. (2017). Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement. *Florida Law Review*, *69*, 181-221. Retrieved from

http://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/uflr69&section=9

Podsakoff, P. M., MacKenzie, S. B. & Podsakoff, N. P. (2016). Recommendations for Creating Better Concept Definitions in the Organizational, Behavioral, and Social Sciences. *Organizational Research Methods*, *19*(2), 159-203. doi: 10.1177/1094428115624965

Rapaport, W. J. (2012). Semiotic Systems, Computers, and the Mind: How Cognition Could Be Computing. *International Journal of Signs and Semiotic Systems (IJSSS)*, *2*(1), 32-71. doi: 10.4018/ijsss.2012010102

Rapaport, W. J. (2017). On the Relation of Computing to the World. In T. M. Powers (ed.), *Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics* (pp. 29-64). New York, NY, USA : Springer International Publishing. doi : 10.1007/978-3-319-61043-6_3

Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B. & Yang, G. Z. (2017). Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, *21*(1), 4-21. doi: 10.1109/JBHI.2016.2636665

Rigdon, E. E. (2016). Choosing PLS path modeling as analytical method in European management research: A realist perspective. *European Management Journal*, *34*(6), 598-605. doi: 10.1016/j.emj.2016.05.006

Robbins, S. & Henschke, A. (2017). Designing For Democracy: Bulk Data and Authoritarianism. *Surveillance & Society*, *15*(3/4), 582-589.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, *115*(3), 211-252. doi: 10.1007/s11263-015-0816-y

Russell, S., Dewey, D. & Tegmark, M. (2015). Research priorities for robust and beneficial artificial intelligence. *AI Magazine*, *36*(4), 105-114. doi: 10.1609/aimag.v36i4.2577

Sabour, S., Frosst, N. & Hinton, G. E. (2017). Dynamic routing between capsules. In I. Guyon, U. Luxburg, S. Bengio (eds.), *31st Conference on Neural Information Processing System (NIPS 2017)in Long Beach, CA, December 04 – 09, 2017* (pp. 3857-3867). La Jolla, CA, USA : Neural Information Processing Systems Foundation, Inc. Retrieved from http://papers.nips.cc/paper/6975-dynamic-routing-between-capsules.pdf

Samson, K. & Kostyszyn, P. (2015). Effects of Cognitive Load on Trusting Behavior – An Experiment Using the Trust Game. *PLOS ONE, 10*(5), e0127680. doi: 10.1371/journal.pone.0127680

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O. & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies! *Journal of Business Research, 69*(10), 3998-4010. doi: 10.1016/j.jbusres.2016.06.007

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85-117. doi: 10.1016/j.neunet.2014.09.003

Schryen, G., Benlian, A., Rowe, F., Gregor, S., Larsen, K., Paré, G., . . . Yasasin, E. (2017). Literature Reviews in IS Research: What Can Be Learnt from the Past and Other Fields? *Communications of the Association for Information Systems, 41*(30), 759-774. doi: 10.17705/1CAIS.04130

Seifert, C., Aamir, A., Balagopalan, A., Jain, D., Sharma, A., Grottel, S. & Gumhold, S. (2017). Visualizations of Deep Neural Networks in Computer Vision: A Survey. In T. Cerquitelli, D. Quercia & F. Pasquale (eds.), *Transparent Data Mining for Big and Small Data* (pp. 123-144). Cham, Switzerland : Springer International Publishing. doi : 10.1007/978-3-319-54024-5_6

Seyfert, R. & Roberge, J. (eds.) (2016). *Algorithmic Cultures: Essays on Meaning, Performance and New Technologies*. New York, NY, USA : Routledge. Retrieved from https://books.google.ca/books?id=_zMlDwAAQBAJ

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3-55. doi: 10.1145/584091.584093

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., . . . Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484-489. doi: 10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., . . . Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354-359. doi: 10.1038/nature24270

Slonim, N. & Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. In N. Belkin, M.-K. Leong & P. Ingwersen (eds.), *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, July 24 - 28, 2000* (pp. 208-2015). New York, NY, USA : ACM. doi: 10.1145/345508.345578

Sotala, K. & Yampolskiy, R. (2017). Responses to the Journey to the Singularity. In V. Callaghan, J. Miller, R. Yampolskiy & S. Armstrong (eds.), *The Technological Singularity: Managing the Journey* (pp. 25-83). Berlin, Heidelberg : Springer Berlin Heidelberg. doi: 10.1007/978-3-662-54033-6_3

Stent, G. S. (1975). Limits to the scientific understanding of man. *Science*, *187*(4181), 1052-1057. doi: 10.1126/science.1114334

Stohl, C., Stohl, M. & Leonardi, P. M. (2016). Digital Age | Managing Opacity: Information Visibility and the Paradox of Transparency in the Digital Age. *International Journal of Communication*, *10*(5), 123-137. Retrieved from http://ijoc.org/index.php/ijoc/article/view/4466

Templier, M. & Paré, G. (2015). A Framework for Guiding and Evaluating Literature Reviews. *Communications of the Association for Information Systems*, *37*(1), 112-137. Retrieved from http://aisel.aisnet.org/cais/vol37/iss1/6

Tishby, N. & Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In Y. Steinberg, R. Zamir & J. Ziv (eds.), *Proceedings of IEEE Information Theory Workshop (ITW) in Jerusalem, Israel, 26 April-1 May, 2015* (pp. 1-5). Washington, DC, USA : IEEE Computer Society. doi: 10.1109/ITW.2015.7133169

Trexler, J. (2008). Social Entrepreneurship as an Algorithm: Is Social Enterprise Sustainable? *Emergence: Complexity and Organization, 10*(3), 65-85.

Turing, A. M. & Copeland, B. J. (2004). *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life plus The Secrets of Enigma*. Oxford, UK : Oxford University Press. Retrieved from https://books.google.ca/books?id=dSUTDAAAQBAJ

Tzeng, F. Y. & Ma, K. L. (2005). Opening the black box - Data driven visualization of neural networks. In C. T. Silva, E. Croller & H. Rushmeier (eds.), *Proceedings of IEEE Visualization in Minneapolis, MN, October 23-28, 2005* (pp. 383-390). Washington, DC USA : IEEE Computer Society. doi: 10.1109/VISUAL.2005.1532820

Valiant, L. G. (2013). *Probably approximately correct : nature's algorithms for learning and prospering in a complex world*. New York, NY, USA : Basic Books. Retrieved from https://books.google.ca/books?id=LBW_dOJ3hoMC

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM, 27*(11), 1134-1142. doi: 10.1145/1968.1972

Vapnik, V. (2000). *The Nature of Statistical Learning Theory* (2nd ed.). New-York, NY, USA : Springer Science & Business Media. Retrieved from https://books.google.ca/books?id=EqgACAAAQBAJ

Vassev, E. (2016). Safe Artificial Intelligence and Formal Methods. In T. Margaria & B. Steffen (eds.), *Leveraging Applications of Formal Methods, Verification and Validation: Foundational Techniques: 7th International Symposium, ISoLA 2016, Imperial, Corfu, Greece, October 10–14, 2016, Proceedings, Part I* (pp. 704-713). Cham, Switzerland : Springer International Publishing. doi : 10.1007/978-3-319-47166-2_49

Veale, M. (2017). *Logics and practices of transparency and opacity in real-world applications of public sector machine learning.* (version 2). Retrieved from https://arxiv.org/abs/1706.09249v2

Vedder, A. & Naudts, L. (2017). Accountability for the use of algorithms in a big data environment. *International Review of Law, Computers & Technology*, *31*(2), 206-224. doi: 10.1080/13600869.2017.1298547

Willson, M. (2017). Algorithms (and the) everyday. *Information, Communication & Society*, *20*(1), 137-150. doi: 10.1080/1369118X.2016.1200645

Woolley, S. C. & Howard, P. N. (2016). Automation, Algorithms, and Politics| Political Communication, Computational Propaganda, and Autonomous Agents — Introduction. *International Journal of Communication*, *10*(19), 4882-4890. Retrieved from http://ijoc.org/index.php/ijoc/article/view/6298

Yanofsky, N. S. (2011). Towards a Definition of an Algorithm. *Journal of Logic and Computation*, *21*(2), 253-286. doi: 10.1093/logcom/exq016

Yu, N., Yu, Z., Gu, F., Li, T., Tian, X. & Pan, Y. (2017). Deep learning in genomic and medical image data analysis: Challenges and approaches. *Journal of Information Processing Systems, 13*(2), 204-214. doi: 10.3745/JIPS.04.0029

Zarsky, T. (2016). The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making. *Science, Technology & Human Values*, *41*(1), 118-132. doi: 10.1177/0162243915605575