INTERPRETABILITY CHALLENGES IN MACHINE LEARNING MODELS

Gabriel Marín Díaz, Ramón A. Carrasco González, Daniel Gómez González

Universidad Complutense de Madrid (Spain)

gabriel.marin@ucm.es; ramoncar@ucm.es; dagomez@estad.ucm.es

ABSTRACT

Decisions based on Machine Learning (ML) algorithms are having an increasingly significant social impact; however, most of these systems are based on black box algorithms, models whose rules are not understandable to humans. On the other hand, different public and private organisations, as well as the scientific community, have recognised the problem of interpretability, focusing on the development of interpretable models (white box) or on methods that allow the explanation of black box models.

The aim of this article is to propose a review of the historical evolution and current state of Machine Learning algorithms, analysing the need for interpretability. In this sense, the challenges of interpretability will be addressed from different points of view: in the field of research, legal, industry and regulatory bodies.

INTRODUCTION

Can machines think? This question was posed by Alan M. Turing (1950) in the middle of the 20th century. The answer to this question is the proposal of the so-called Turing test. In this test, Artificial Intelligence (AI) is considered to be the way of acting that imitates the intelligent behaviour of human beings. From then until now, AI has been surpassing human beings in tasks for which intelligence was supposed to be required: strategy games such as chess, driving vehicles, composing symphonies, automatic planning, and a long etcetera that seems to have no end in sight. In fact, the changes that have taken place in recent decades in the telecommunications sector, accompanied by the development of information storage and processing capacity, have led to a paradigm shift that has been given the name of Industry 4.0.

AI corresponds to a field of knowledge that includes Machine Learning (ML) and Deep Learning (DL). In both fields, to solve a problem, models are trained to learn the problem in question from existing data. Once the rules are obtained, we can apply them to new data sets to produce the appropriate answers by applying the rules learned from experience. To perform ML processes, at least three fundamental parts are necessary: input data, the expected results and the measurement of the algorithm's performance so that the algorithm's work can be adjusted through feedback processes (Casella et al., 2013).

An ML model, once implemented, can complete a task much faster and more reliably than any human, delivers consistent results "reliably" and can be infinitely replicated. Training a person to perform a task with the same efficiency is costly and can take years.

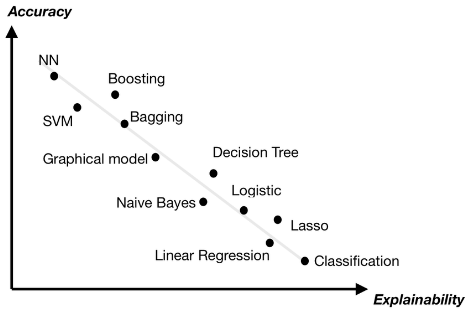
An important aspect of the use of ML is the interpretability of the models once they have been trained. From this point of view some authors distinguish two types of models (Liu et al., 2016):

White-box models are models whose predictive or pattern-identifying behaviour can be clearly explained based on the variables involved.

Black box models are models whose rules are not understandable in a simple way for humans, it would be very difficult to explain how the system came to a certain decision on a certain input. For example, Artificial Neural Networks (ANNs) and DL algorithms in general, such that millions of operations are needed to describe a deep neural network, and there is no way to understand the model in its entirety, hiding how the machine solves a task in increasingly complex models. (Liu et al., 2016).

Some authors even question the interpretability of white-box algorithms (Z. C. Lipton, 2018). Figure 1 shows, as a general rule, that the higher the interpretability of the ML algorithm, the lower its degree of flexibility and consequently the lower its degree of reliability. In other words, there is currently no doubt that the most powerful algorithms are not interpretable.

Figure 1. Interpretability vs Flexibility of ML algorithms.



Source: (Duval, 2019)

In low-risk environments it may not be relevant to understand why a decision has been taken. However, on most occasions, human beings should understand why a decision that affects them individually or collectively has been made. Examples include: a loan decision, a medical decision, self-driving cars, a selection process for a particular job...

In a study conducted in 2019 by Brandon Fornwalt, Geisinger Medical Center, Pennsylvania, they trained two AI algorithms capable of predicting the risk of death in the first year by reading electrocardiograms, even from apparently "normal" people; the algorithm's accuracy was 85% (Samad et al., 2019).

It has been shown that ML models learn very well from the data, but they also pick up biases that may be built into the training data voluntarily or unintentionally. This can make the training model potentially sectarian and discriminate against certain groups and individuals. These potential biases are a key point in investigating the problem of interpretability (Miller, 2019).

In this paper we will review the challenges facing the problem of interpretability in ML models, according to the following structure: in section 2 we will trace the historical evolution of interpretability models, in section 3 we will address the importance of decision-making in this context and how interpretability is a determining factor in the trustworthiness of ML models, and in section 4 we will address the challenges in the field of research, legal, industry and regulatory bodies. In section 5 we will review the interpretability indicators, identifying the quantitative and qualitative factors that make an ML model interpretable, and finally we will draw conclusions.

HISTORICAL DEVELOPMENT OF INTERPRETABILITY

From a historical point of view, in 1950 Turing created the test that bears his name, in 1952 Arthur L. Samuel created the first algorithm capable of learning, in 1956 the concept of Artificial Intelligence was born, in the 70s pattern recognition algorithms emerged, in the 80s expert systems based on rules appeared, the concept of ML began to gain relevance in the 90s being currently one of the most popular subfields within AI, closely linked to mathematical statistics.

Historically, the focus of AI research has shifted towards the implementation of algorithms and models focusing on predictive power to the detriment of interpretability. Model interpretability was emphasised in early machine learning research. The 1970s and 1990s saw the emergence of initiatives such as MYCIN (Britannica, 2018), GUIDON (Clancey, 1987). From the 1980s to the 1990s, systems for tracking alternative lines of reasoning (TMS) were developed. In the 1990s, initiatives emerged in the context of explaining neural networks in healthcare. In 2010, concerns about bias in AI decision-making led to a demand for transparent artificial intelligence and a focus on the interpretability of ML models.

In addition, during recent years we have seen the expansion of social networking systems, which are underpinned by the speed of information processing, communications and storage capacity. As a consequence, due to the exponential increase in heterogeneous data collection and the enormous amount of computational power, machine learning (ML) systems are present in our lives, achieving higher predictive performance and, for most of them, greater complexity (Carvalho et al., 2019).

In practice, what we want is for algorithms to be explainable, i.e. that their operations can be understood by human beings. Despite the correspondence between the two terms, interpretable vs. explainable, there are authors who develop a certain differentiation between the two concepts (Rudin, 2019). Initiatives such as Explainable Artificial Intelligence (XAI) (Gunning et al., 2019), focusing on the interpretability of machine learning algorithms, aims to move towards an interpretable AI model.

On the other hand, and concerning the characteristics that should be attached to interpretable models (Molnar, 2019), we can highlight: the explanations should be contrasting (P. Lipton, 1990), the question we ask ourselves is why a certain prediction was made rather than another, we need to understand by comparison. Furthermore, explanations are selected, i.e., from the set of causes that can give a certain explanation, we are used to selecting one or two causes as the ones most linked to the explanation. Explanations are social, they are part of an interaction between the explainer and the receiver of the explanation where in many cases the social environment is involved. Explanations focus on the abnormal (Kahneman, 1981), causes that are attributed with high potential but low probability. Explanations are true, good explanations prove to be true, the event should be predicted with the highest possible probability. Explanations are consistent with prior beliefs, this is called confirmation bias, devaluing explanations that do not match your beliefs (Nickerson, 1998). Good explanations are general and probable, in the absence of an abnormal scenario, general causes are good explanations (Gaussian curve).

As can be seen, the concept of interpretability suggests the involvement of more than one area of knowledge (Carvalho et al., 2019), at least three stand out: data science developing predictive models, social sciences leading to understanding, and human-machine interaction to empower the user (Abdul et al., 2018).

Interpretability is therefore a necessary milestone for the success of ML and AI itself. As stated by (Roy, 2017), "In the end, mathematical models should be our tools, not our masters", which is only possible with interpretability.

Interpretation methods for machine learning can be classified according to several criteria (Z. C. Lipton, 2018).

Intrinsic or post hoc? This criterion distinguishes whether interpretability is achieved by restricting the complexity of the machine learning model (intrinsic) or by applying methods that analyse the model after training (post hoc). In the former case, interpretability is inherent to the model, in the latter case, the methods may or may not be decoupled from the ML model.

Especific or agnostic? Interpretation tools are limited to specific model classes e.g. linear regression, or applicable to any model once trained (post hoc). Agnostic tools can be used on any machine learning model and separate the explanation from the type of model. They offer the freedom to choose a set of models to address a problem and then compare them.

Local or global? Does the interpretation method explain an individual prediction or the entire behaviour of the model? Or is the scope somewhere in between?

IMPORTANCE OF INTERPRETABILITY IN ML

According to Miller "Interpretability is the degree to which a human can understand the cause of a decision" (Miller, 2019). This means that in the interpretation of a model there is a directly proportional link to causality, understanding why a prediction was made by the model.

On the other hand, a correct prediction only partially solves the original problem, an explanatory black box model that has 85% agreement with the original model does indeed explain the original model most of the time, however, it is wrong 15% of the time. Therefore, confidence in this black box model is limited to that 85% reliability. The higher the interpretability of a machine learning model, the easier it is for someone to understand why certain decisions or predictions have been made. In some cases it may not be relevant to understand why a certain decision has been made, especially in a low risk environment (Doshi-Velez & Kim, 2017). In most cases, however, we need to understand why, as it can help us to better understand the problem and the reasons why a model may fail.

The trend in recent years has been to take advantage of the characteristics of ML and in particular the predictive power of black box algorithms for high-risk decision making, for example in legal, financial and health services, all of which have a profound impact on society and in particular on human lives (Rudin, 2019), a fundamental point in the investigation of the problem of its interpretability (Miller, 2019; Molnar, 2019).

The following are some examples of biases applied by ML algorithms in different technologies and areas:

On 7 November 2019, Ruby on Rails creator and entrepreneur David Heinemeier Hansson shared a disturbing story on Twitter (Business, 2019), alleging that the Apple card was discriminating against his wife. Both he and his wife applied for this card, but he received a credit limit 20 times higher than she did, even though they applied for the card at the same time, and filed joint tax returns.

In the autumn of 2019, Google unveiled a Machine Learning technology called BERT (Bert, 2019) to improve its search criteria by incorporating the context of words accompanying the search object. However, the data it works with corresponds to the largest digital library in history, bringing with it decades of biases and prejudices that are built into the search algorithm, and are likely to be perpetuated.

In 2015, the Google Photos app labelled two African-Americans as "gorillas". (BBC Mundo Tecnología, 2015). Google engineers analysed the account and discovered that the algorithm had problems adjusting for photo contrast, lighting and skin tone. In addition, they confessed that, due to this same problem, the algorithm labelled white-skinned people as dogs and seals.

In 2016, some of LinkedIn's algorithms were found to have a gender bias (Day, 2016), recommending better paid jobs for men. This casuistry may be reinforced by the fact that high-paying jobs are predominantly held by men.

In 2016 Microsoft launched "Tay" (BBC Mundo, 2016), a chatbot whose purpose was to mimic the behaviour of a curious teenager seeking to engage in casual conversation on social media with a target audience of 18-24 year olds. In less than 24 hours, Tay, through tweets, showed her empathy for Hitler or her support for genocide by answering questions from social media users.

In 2016, the COMPAS algorithm (Correctional Offender Management Profiling for Alternative Sanctions) to predict recidivism, developed by Northpointe (now Equivant), was accused of bias against African Americans (Larson et al., 2016).

In 2018, oncologists criticised IBM's Watson for Oncology for providing unsafe and inaccurate recommendations (Ross, 2018).

In 2018, Amazon's resume screening system was found to be biased against women (Dastin, 2018).

In 2019, algorithms behind Apple's credit card are accused of gender bias (Fast Company, 2019).

As can be seen, there are a significant number of predictive models whose outcome determines a negative impact on people's safety and rights, leading to serious violations of ethical and equity principles. In this context, it is essential to build tools that allow for model exploration, in particular to explain the model, examine and evaluate its performance, and understand its weaknesses and failures.

Equally important are algorithmic audits to detect discrimination and bias, and to incorporate ethical values into these systems (Carvalho et al., 2019).

According to (Doshi-Velez & Kim, 2017) the aspects that could be optimised through interpretability are as follows:

Impartiality, unbiased, non-discriminatory predictions.

Privacy, protection of information.

Reliability, small changes in the data input do not affect the prediction.

Causality, only collect causal relationships (cause-effect).

Confindence, systems must explain their decisions in order to be reliable.

The fundamental objective is to gain trust and social acceptance of ML algorithms through interpretability.

INITIATIVES TOWARDS INTERPRETABLE AI

Currently, there is no real consensus on what interpretability in Machine Learning models is, nor is it clear how to measure interpretability. However, technology companies, international organisations and public administrations are aware of the problem and are taking steps to mitigate the consequences of discriminatory bias in algorithms.

Technology Companies

IBM launched the Fairness 360 Kit project in 2018 (Hughes et al., 2020), this open source toolkit helps to examine, report and mitigate discrimination and bias in machine learning models. Adversarial Robustness 360 (ART) Toolbox (Adversarial Robustness Toolbox, 2021) is a Python library for machine learning security. AIX360 Toolkit (AI Explanability 360, 2021) is a comprehensive open source toolkit with various algorithms, codes, guides, tutorials and demos that support the interpretability of machine learning models.

Microsoft has a model interpretation SDK in Azure Machine Learning for use in Python (Microsoft, 2021).

Google has an API, Explainable AI (Hughes et al., 2020), is a set of tools and frameworks capable of helping to debug and understand the behaviour of Machine Learning models.

H2O Driverless AI, a machine learning platform (H2O.ai., 2020) offered by H2O.ai, offers interpretability as one of its distinguishing features.

DataRobot (DataRobot, 2021), is another commercialised ML solution, "includes several components that result in models that are highly interpretable by humans".

Google Vizier is a service for optimising black box models. (Golovin et al., 2017).

Facebook, in collaboration with Georgia Tech, published an article showing a tool for visual exploration of industrial-scale DNN models (Kahng et al., 2018).

Uber recently announced Manifold, a model-agnostic visual debugging tool for ML (Zhang et al., 2018).

Other companies are taking steps in the same direction; however, perfection cannot be expected, there will always be undetected biases, or biases that cannot be eliminated.

Legislation, Organisations and Regulatory Documents

As business and government decisions become increasingly automated, the need to protect against black box algorithms will be critical. We will need to know how and why decisions are made, understanding is crucial to move forward safely. To this end, it will be necessary to work on the control and auditing of algorithms whose decisions directly affect people, independent bodies will be needed that are capable of determining the "quality" of the algorithm, providing sufficient guarantees to citizens, thus increasing the social acceptance of this type of practice. Ensuring that the following qualities are met: fairness, privacy, reliability, robustness, causality, trustworthiness (Doshi-Velez & Kim, 2017).

Profiling and automated decisions can pose significant risks to individual rights and freedoms. European and Spanish data protection legislation obliges and requires certain safeguards. Article 22 of the GDPR (UE, 2016) provides that European citizens have the right not to be subject to a decision based solely on automated means, including profiling, if the decision produces legal effects which significantly affect them in a similar way.

On the other hand, we have the ISO/IEC 27001 standard (Blackmer, 2018), which aims to ensure the confidentiality, integrity and availability of an organisation's information and the systems and applications that process it, this standard has been developed by the International Organisation for Standardisation (ISO), and will have to adapt its content to the needs arising from the interpretability in the IA (Weller, 2019).

One of the most notable entities in the field of AI research, Defense Advanced Research Projects (DARPA), created the XAI programme (Gunning et al., 2019). In 2016 the White House Office of Science and Technology Policy (OSTP) published the US report on AI entitled "Preparing for the Future of Artificial Intelligence". (Bundy, 2017).

The Royal Society, which is the UK Academy of Sciences, published a report on its machine learning project in April 2017 (Royal Society of Great Britain, 2017).

In Spain, the technical subcommittee for standardisation CTN 71/SC 42 - Artificial Intelligence and Big Data was set up in December 2019 (UNE, 2021) precisely to elaborate standards in the field of AI, participating in the development of global standards being developed in the international committee ISO/IEC JTC 1/SC 42 Artificial Intelligence (Standardization, 2021).

In April 2018, the European Commission published the following communication on Artificial Intelligence for Europe (Commission, 2018). In 2019, the High-Level Expert Group on Artificial Intelligence formulated guidelines on trustworthy AI (European Commission, 2019). In parallel, the first coordinated plan on AI was published in December 2018 as a joint commitment with the Member States (Digitales et al., 2020).

The Commission's White Paper on AI, published in 2020, sets out a clear vision for AI in Europe: "an ecosystem of excellence and trust that lays the foundation for today's proposition" (Comisión Europea, 2020).

In April 2021, the European Commission, in coordination with member states and with the aim of strengthening trust and excellence in AI, launched a risk-based approach that penalises, and even bans, AI systems that are considered a clear threat to security. High-risk systems will be subject to strict obligations before they can be placed on the market (Munchen, 2021).

Gradually, both EU and non-EU countries will join such initiatives, so as to offer a glimmer of hope and try to ensure reliability in AI systems.

Science

The easiest way to achieve interpretability is to use interpretable ML algorithms (white box models), including linear regression, logistic regression, decision trees, RuleFit and Naive Bayes. (Molnar, 2019). From these models, features can be extracted in terms of other features that allow the model to be defined and interpreted at a global level (Sundararajan et al., 2017).

On the other hand, there are model-specific methods of explanation, many of which are designed to be used with neural networks that are difficult to interpret (black box models). Another option is to extract knowledge from a more complex model by approximating it with an interpretable model (Bastani et al., 2017; Tan et al., 2018).

Finally, we have the agnostic methods of explanation, which do not depend on the ML model, and are post hoc, the great advantage of these models over the specific ones is their flexibility.

An overview of agnostic models is represented in the table 1 (Carvalho et al., 2019):

The current trend is to focus on model-independent interpretation tools; it, is much easier to automate interpretability if we separate the interpretation method from the model used. With agnostic methods we can replace both the learning model and the interpretation method, the capabilities provided by this system are highly scalable. (Carvalho et al., 2019; Ribeiro et al., 2016; Molnar, 2019).

Explanation Method	Scope	Result
Partial Dependence Plot	Global	Feature Summary
Individual Condition Expectation	Global / Local	Feature Summary
Accumulated Local Effects Plot	Global	Feature Summary
Feature Interaction	Global	Feature Summary
Feature Importance	Global / Local	Feature Summary
Local Surrogate Model	Local	Surrogate Interpretable Model
Shapley Values	Local	Feature Summary
BreakDown	Local	Feature Summary
Anchors	Local	Feature Summary
Counterfactual Explanations	Local	(new) Data Point
Prototypes and Criticisms	Global	(existent) Data Point
Influence Functions	Global / Local	(existent) Data Point

As we have seen in this document, interpretability is not only a scientific question, other areas of knowledge linked to the human being are involved. The need for interpretability is inherent to the desire to know, to the human being's need to answer the question of why.

INTERPRETABILITY INDICATORS

Can we measure and evaluate interpretability? Despite all the work being done in different areas of knowledge, this question unfortunately remains unanswered. However, the work that is being done is oriented along two clear lines: the use of ML algorithms that allow a high degree of precision and making the decision adopted by these systems interpretable, explainable to human beings.

A review of the literature suggests that little work has been done to develop models to measure and evaluate interpretations, so that the most appropriate explanation can be chosen (Honegger, 2018). However, we can distinguish between two types of indicators when comparing and evaluating explanations (Carvalho et al., 2019), quantitative and qualitative.

Among the qualitative indicators (Doshi-Velez & Kim, 2018) suggests the following questions:

What are explanations composed of? Which features are predominant in an explanation?

How many subsets of blocks of features can an explanation be made up of, and if we remove any blocks, is the result affected?

How are these blocks formed? What composition should be given between the blocks?

What relationships might be more intuitive to humans?

Are any random processes part of the explanation?

For quantitative indicators (Sundararajan et al., 2017), (Honegger, 2018) established a framework for measuring the consistency of explanatory methods whose prediction must be consistent with human explanation. It is necessary according to (Honegger, 2018) to relate the object (instance and prediction) to its subsequent explanation (importance value of features).

Identity. Identical objects must have identical explanations. If a method of explanation is asked to explain a certain object, the explanations it gives must be the same.

Separability. Non-identical objects cannot have identical explanations. It follows from the previous premise.

Stability. Similar objects must have similar explanations. If slight perturbations considerably modify the response, the system is not stable.

In addition, other variables must be considered, such as completeness, the audience has to verify the validity of the explanation. Correctness, the explanation must generate confidence. And finally, compactness, the explanation must be precise, brief, and concise.

CONCLUSIONS

And at company and individual level, what can be done? We propose to intervene on the following aspects.

Addressing biases

While, as we have said, the task will not be easy, we do not know how black box algorithms work, but we can act on the biases so that the decisions taken by the algorithms are aligned with the rights of the people.

Digital maturity

Machine learning is the subject of many expectations, but are companies ready for data governance? Science is constantly developing machine learning tools, but can they be integrated into a company's business processes? Most companies have grown based on technological silos, integrating pieces with little or no scalability. There is no such thing as a single piece of data; the fundamental task of data scientists is to "find out" where the information is to be able to analyse and make predictive models. Their core business has more to do with an Agatha Christie novel than with analysis and predictive modelling useful for the business. The expectations generated by the media and the occasional guru about AI and its application are unlikely to be fulfilled until the business culture of the short-term changes.

Machine learning will grow, not at the speed it is touted, but slowly and steadily. Fundamental to this is the process of business digitalisation that starts from a single data model, from the integration of all the company's information to be able to extract working models where AI can develop. Better formulas are needed to integrate AI into the business processes of companies, perhaps it would be useful to develop machine learning tools that are easy to use and can be automatically integrated with business management processes, this would help to make a technological leap in the digitisation process and will be the first step from childhood to youth.

The next step towards maturity could be the adoption of a full automation model of business management processes, tasks could be posed as decision problems solved by machine learning.

Interpretability as a catalyst

At this point, interpretability will be critical to ensure that algorithms are responsive to reality, trying to minimise the impact of biases. On the other hand, transparency is the norm in any organisation - decisions need to be supported by an understanding of the underlying tasks. The interpretability of algorithms must be fundamental to trust in predictive black box models. If we use interpretation-agnostic methods, we can automatically apply them to any model that emerges in a machine learning process and train surrogate models that improve the predictions. If we are able to do this, we may be able to improve our understanding of intelligence and become better at creating intelligent machines.

The opportunities are obvious, but so are the associated dangers. In an increasingly anumerical society (Hand & Paulos, 1992) where decision making is usually done through System 1 thinking: quick, intuitive and emotional (Kahneman, 2012), versus System 2 thinking: slow, deliberative and logical. We can sense that this speed, immediacy of the everyday and short-termism can invade us, leaving the decisions that require thought and meditation to a third party (the machine).

There is room for improvement and progress, the expectations are good, but so are the challenges!!!!

KEYWORDS: Machine Learning, Interpretability, Deep Learning, Bias, Artificial Intelligence.

REFERENCES

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). 17. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*. https://doi.org/10.1145/3173574.3174156

Adversarial Robustness Toolbox. (2021). *Adversarial Robustness Toolbox*. https://adversarial-robustness-toolbox.org/

AI Explanability 360. (2021). *AI Explanability 360*. https://aix360.mybluemix.net/

Bastani, O., Kim, C., & Bastani, H. (2017). 137. Interpreting blackbox models via model extraction. *ArXiv*.

BBC Mundo. (2016). *Tay, la robot racista y xenófoba de Microsoft*. Bbc. https://www.bbc.com/mundo/noticias/2016/03/160325_tecnologia_microsoft_tay_bot_adolescente_inteligencia_artificial_racista_xenofoba_lb%0Ahttp://www.bbc.com/mundo/noticias/2016/03/160325_tecnologia_microsoft_tay_bot_adolescente_inteligencia_artificial_raci

BBC Mundo Tecnología. (2015). *Google pide perdón por confundir a una pareja negra con gorilas*. Bbc. https://www.bbc.com/mundo/noticias/2015/07/150702_tecnologia_google_perdon_confundir_afroamericanos_gorilas_lv

Bert, G. (2018). *Google BERT*. https://cloud.google.com/tpu/docs/tutorials/bert

Blackmer, W. S. (2018). 84. EU general data protection regulation. *American Fuel and Petrochemical Manufacturers, AFPM - Labor Relations/Human Resources Conference 2018*, *2014*(April), 45–62. https://doi.org/10.1308/rcsfdj.2018.54

Britannica, E. (2018). *MYCIN*. https://www.britannica.com/technology/MYCIN

Bundy, A. (2017). 20. Preparing for the future of Artificial Intelligence. *Ai & Society*, *32*(2), 285–287. https://doi.org/10.1007/s00146-016-0685-0

Business, C. (2019). *Apple co-founder Steve Wozniak says Apple Card discriminated against his wife*. https://edition.cnn.com/2019/11/10/business/goldman-sachs-apple-card-discrimination/index.html

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). CAT. A - Machine learning interpretability: A survey on methods and metrics. *Electronics (Switzerland)*, *8*(8), 1–34. https://doi.org/10.3390/electronics8080832

Casella, G., Fienberg, S., & Olkin, I. (2013). An Introduction to Statistical Learning. In *Springer Texts in Statistics*. http://books.google.com/books?id=9tv0taI8l6YC

Clancey, W. J. (1987). The GUIDON Program. *MIT Press Series in Artificial Intelligence*.

Comisión Europea. (2020). Libro Blanco sobre la Inteligencia Artificial - un enfoque europeo orientado a la excelencia y la confianza. *Comisión Europea*, 1–31. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_es.pdf

Commission, E. (2018). *Artificial Intelligence for Europe - Communication*. https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF

Dastin, J. (2005). *Amazon scraps secret AI recruiting tool that showed bias against women*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

DataRobot. (2021). *DataRobot*. https://www.datarobot.com/wiki/interpretability/

Day, M. (2016). *How LinkedIn's search engine may reflect a gender bias*. The Seattle Times. https://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias/

Digitales, S., Unidos, E., Europa, H., & Digital, P. E. (2020). *Los Estados miembros y la Comisión colaborarán para impulsar la inteligencia artificial « fabricada en Europa » Contexto Más información*. 2019–2021.

Doshi-Velez, F., & Kim, B. (2017). *41. Towards A Rigorous Science of Interpretable Machine Learning*. *Ml*, 1–13. http://arxiv.org/abs/1702.08608

Doshi-Velez, F., & Kim, B. (2018). *152. Considerations for Evaluation and Generalization in Interpretable Machine Learning*. 3–17. https://doi.org/10.1007/978-3-319-98131-4_1

Duval, A. (2019). *Explainable Artificial Intelligence (XAI) Explainable Artificial*. *April*. https://doi.org/10.13140/RG.2.2.24722.09929

European Commission. (2019). *COM(2019) 168 final Building Trust in Human Centric Artificial Intelligence*. 11. https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

Fast Company. (2019). *I applied for an Apple Card. What they offered was a sexist insult*. https://www.fastcompany.com/90429224/i-applied-for-an-apple-card-what-they-offered-was-a-sexist-insult

Golovin, D., Solnik, B., Moitra, S., Kochanski, G., Karro, J., & Sculley, D. (2017). 8. Google vizier: A service for black-box optimization. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *Part F1296*, 1487–1496. https://doi.org/10.1145/3097983.3098043

Goodman, B., & Flaxman, S. (2017). 88. European union regulations on algorithmic decision making and a "right to explanation." *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). 18. XAI-Explainable artificial intelligence. *Science Robotics*, *4*(37), 0–1. https://doi.org/10.1126/scirobotics.aay7120

H2O.ai. (2020). *H2O Driverless AI*. https://www.h2o.ai/products/h2o-driverless-ai/

Hand, D., & Paulos, J. A. (1992). Innumeracy: Mathematical Illiteracy and its Consequences. In *Applied Statistics* (Vol. 41, Issue 1). https://doi.org/10.2307/2347643

Honegger, M. R. (2018). *79. Shedding Light on Black Box Machine Learning Algorithms*. *August*.

Hughes, R., Edmond, C., Wells, L., Glencross, M., Zhu, L., & Bednarz, T. (2020). *eXplainable AI (XAI)*. 1–62. https://doi.org/10.1145/3415263.3419166

Kahneman, D. (1981). *The Simulation Heuristic*.

Kahneman, D. (2012). Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2011. In *Etc* (Issue October).

Kahng, M., Andrews, P. Y., Kalro, A., & Chau, D. H. P. (2018). 39. ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models. *IEEE Transactions on Visualization and Computer Graphics*, *24*(1), 88–97. https://doi.org/10.1109/TVCG.2017.2744718

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How We Analyzed the COMPAS Recidivism Algorithm*. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Lipton, P. (1990). Contrastive explanation. *Contrastivism in Philosophy*, 11–34. https://doi.org/10.4324/9780203117477

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 35–43. https://doi.org/10.1145/3233231

Liu, H., Cocea, M., & Gegov, A. (2016). Interpretability of computational models for sentiment analysis. *Studies in Computational Intelligence*, *639*(March), 199–220. https://doi.org/10.1007/978-3-319-30319-2_9

Microsoft. (2021). *Instalar el SDK de Azure Machine Learning para Python*. https://docs.microsoft.com/es-es/python/api/overview/azure/ml/install?preserve-view=true&view=azure-ml-py

Miller, T. (2019). 95. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Molnar, C. (2019). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. *Book*, 247. https://christophm.github.io/interpretable-ml-book

Munchen, T. U. (2021). European approach to Artificial Intelligence. *E-Conversion - Proposal for a Cluster of Excellence*, 29–50. https://ec.europa.eu/digital-single-market/en/news/communication-building-trust-human-centric-artificial-intelligence

Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Zeitschrift Für Neurologie*, *199*(1–2), 145–150. https://doi.org/10.1007/BF00316552

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *Model-Agnostic Interpretability of Machine Learning*. *Whi*. http://arxiv.org/abs/1606.05386

Ross, C. (2018). Watson for Oncology. *STAT*, 1–30. papers3://publication/uuid/5566F158-417A-46D3-B583-04EE273812A1

Roy, M. (2017). 80. Cathy O'Neil. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishers, 2016. 272p. Hardcover, $26 (ISBN 978-0553418811). In *College & Research Libraries* (Vol. 78, Issue 3). https://doi.org/10.5860/crl.78.3.403

Royal Society of Great Britain. (2017). 24. Machine learning : the power and promise of computers that learn by example. In *Report by the Royal Society* (Vol. 66, Issue January).

Rudin, C. (2019). 9. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Samad, M. D., Ulloa, A., Wehner, G. J., Jing, L., Hartzel, D., Good, C. W., Williams, B. A., Haggerty, C. M., & Fornwalt, B. K. (2019). Predicting Survival From Large Echocardiography and Electronic Health Record Datasets: Optimization With Machine Learning. *JACC: Cardiovascular Imaging*, *12*(4), 681–689. https://doi.org/10.1016/j.jcmg.2018.04.026

Standardization, I. O. (2021). *ISO*. International Organization for Standardization. https://www.iso.org/committee/6794475.html

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *ArXiv*.

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018). 77. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 303–310. https://doi.org/10.1145/3278721.3278725

UE. (2016). *Artículo 22 UE RGDP*. https://www.privacy-regulation.eu/es/22.htm

UNE. (2021). *UNE Normalización Española*. https://www.une.org/encuentra-tu-norma/comites-tecnicos-de-normalizacion/comite/?c=CTN 71/SC 42

Weller, A. (2019). 85. Transparency: Motivations and Challenges. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11700 LNCS*(Section 2), 23–40. https://doi.org/10.1007/978-3-030-28954-6_2

Zhang, J., Wang, Y., Molino, P., Li, L., & Ebert, D. S. (2018). 40. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *ArXiv*, *25*(1), 364–373.