Analysis of Trustworthiness in Machine Learning and Deep Learning

Mohamed Kentour, Joan Lu
University of Huddersfield
Huddersfield, UK
email: Mohamed.kentour@hud.ac.uk, J.Lu@hud.ac.uk

***Abstract*—Trustworthy Machine Learning (TML) represents a set of mechanisms and explainable layers, which enrich the learning model in order to be clear, understood, thus trusted by users. A literature review has been conducted in this paper to provide a comprehensive analysis on TML perception. A quantitative study accompanied with qualitative observations have been discussed by categorizing machine learning algorithms and emphasising deep learning ones, the latter models have achieved very high performance as real-world function approximators (e.g., natural language and signal processing, robotics, etc.). However, to be fully adapted by humans, a level of transparency needs to be guaranteed which makes the task harder regarding recent techniques (e.g., fully connected layers in neural networks, dynamic bias, parallelism, etc.). The paper covered both academics and practitioners works, some promising results have been covered, the goal is a high trade-off transparency/accuracy achievement towards a reliable learning approach.**

Keywords—Trustworthy machine learning; deep learning; transparency/accuracy; perception.

I. INTRODUCTION

A lot of research flows and advanced computing techniques are inspired by machine learning [1], this multi-disciplinary area merges the human understanding with machine physical capabilities in order to retrieve meaningful correlations and to improve computation. With the tremendous data deluge [2], the users became unable to analyse the amount of data without a machine intervention, due to the high processing power and their precision which became a paramount. For instance, in medical domain, surgical robots (i.e., endoscopic robot for brain surgical) make critical decisions on patients' life [3]; autopilot systems share a critical part of security control with human pilots [4]; space missions become more reliable and faults tolerant [5], etc.

However, for achieving a reasonable and optimal outcome; a user needs to be confident about the decisions made by these learning systems which may include his perception about both the intelligent model and his own knowledge, this is qualified as trust design modeling [6]. From an expert perspective, these learning models' outcomes could be understood and interpreted. For example, by using visualization analytics through an interactive model [7]. But, for a naive user (i.e., how children relate to robots) this may be quite misunderstood. By this research, we aim to extract reliable metrics that most impact users' trust with the learning-based systems; this will be done by analysing practical models (e.g., IBM 360°, DARPA, etc., (see B)) and addressing some of their limits. Furthermore, we investigate a model decomposition that helps include those contextual metrics in to the learning process.

The rest of this paper is organised as follows: Section II qualifies trust in Machine Learning (ML) by covering the main approaches and describing the techniques and results via quantitative and qualitative way. Section III depicts a brief evaluation of this work through a comparative analysis with recent surveys; this is followed by an emphasis on our analysis' contribution and possible answers to the defined research questions. Section IV highlights a critical view of the previous approaches by emphasising some gaps. Ethics related to trust in ML were identified in section V. Section VI concludes and gives some potential research directions.

A. Research questions

The following viewpoints are proposed to frame the present research:

- Trustworthiness towards users' confidence.
- Data-driven approach to interpret ML algorithms.
- Metrics in order to explain ML/Deep Learning (DL) predictions.
- context: academic and industrial projects.

These boundaries were developed by the following questions:

- what are the dimensions of trust in ML?
- Does the inner ML mechanism impact users' reactions?
- How can data-driven metrics bridge learning processes with human understanding compared to explainable AI (XAI) approaches?
- Which ML models (clustering, neural-nets, etc.) are most targeted and/or suitable for transparency?
- Are current research flows more data-driven or XAI inspired, and what impact do they have on practitioners?

B. Journal paper selection

Three main research databases have been invoked in order to retrieve the discussed papers from journals with reference to trustworthy machine learning. First, ScienceDirect has been queried to extract research/review articles with a reference to explainable and trust in machine learning. Then, the ones referring to explainable and trust in deep learning have been extracted using Springer database. After that, Results had to be refined (see TABLE I) to exclude the records which are not user-centered ones: first, by expanding the research up to the explainable AI (n = 165 articles); second, by executing 'AND' between previously mentioned articles (n = 73 article).

ISBN: 978-1-61208-865-5

TABLE I

RESEARCH DATABASES AND MOST RELATED SUBJECTS.

Research database	Key-word	Number of journal papers	Subject
Springer	explainable trust deep learning	84	Compute Science and Artificial Intelligence
ScienceDirect	explainable trust machine learning	35	Computers and Security
ACM Digital Library	User centered explainable Artificial Intelligence	165	Explainable Artificial Intelligence

II. BACKGROUND AND LITERATURE REVIEW

Nowadays, machine Learning (ML) dominates several domains: business, finance, industry, travel, psychology, medicine, etc. ML-systems are now seen as a black-box because of advanced data driven techniques [8] that hides the way how decisions are made, from here the notion of trust (or trustworthiness) arises and becomes crucial [9]. As trust within ML is a general term (i.e., ethics, certifications, privacy, etc.), techniques to include end-user's perception within the learning process are not well covered [10]. To this end, it has been decided to approach trust from transparency in ML, as this is an emergent field in ML and commonly investigated within a trade-off performance and transparency. Therefore, a user-centered investigation around trustworthiness in ML-based systems has been conducted in this paper; we end by a model decomposition (see C) as a method to include users' perception into the learning flow.

A. Interpretable qualification of trust in machine learning

To qualify trust for learning systems some challenges have been addressed regarding users' interaction (i.e., design complexity, hidden layers in fully automated systems [11], users' behaviour and beliefs, etc.). Those arguments justify a modern vision of trust in smart networking protocols in accordance with the emergence of cloud computing and machines' internal architecture improvement [11], where a selective smart agent (human simulation) is involved to pick the service resource among several nodes (options). Figure 1 illustrates the main constructs of trust in intelligent systems and their variations.

Users' confidence which depicts trustworthiness has a strong dependency with both user' behaviour and intention which are together fundamental to approach users from ML systems and improve interactivity, this manner to tackle trust is called data driven (Interpretable) approach.

An interesting study [12] which aimed to increase trust between buyers and sellers for an e-marketplace by using visual stereotyping, results show accurate measures on limited knowledge. This work has been recently extended [13], [14] where sensor devices have been developed to capture user's profiles and interpret their intentions.

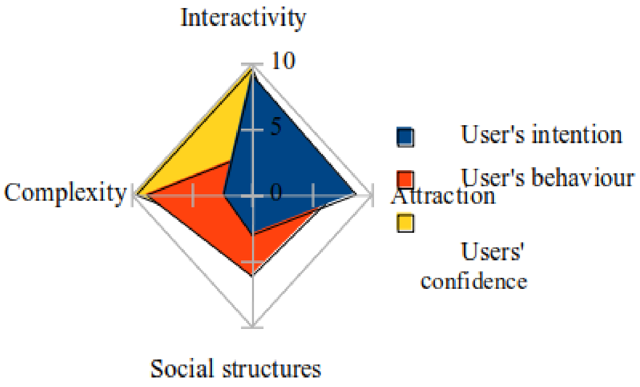

Figure 1. Formalism of Trustworthiness in computing environments.

The main challenges in this area is how to bridge qualitative and quantitative measurements to fit with the learning model [15]. In [16], interpretation metrics have been proposed (i.e., replicability) to evaluate learning predictions and measure the effect of ML decision on people behaviour. This may be seen as an extension of the model (see Figure 1). Through an interactive process (human-machine or human-human), [17] have proposed an incremental model to give an in-depth interpretation of ML model by going through real-world scenarios and distinguishing simple, reflective and pragmative trust.

- Techniques and Results

The following TABLE II highlights some works on interpretable ML.

TABLE II

WORKS ON INTERPRETABLE TRUST IN ML.

Authors	Data type	Techniques	Results
[15]	Application dependent dataset	Data driven technique applied on a matrix of: real cases' rows and learning methods' columns F(knowledge, methods).	Relevant separation of interpretable definitions and evaluation based on the background knowledge and application specifications.
[13]	Limited data (numeric and nominal).	Fuzz System (for numeric data) and semantic process (ontology) for nominal attributes applied on a Decision Tree model (FSDT).	Better results shown with all data partition compared to each technique applied separately
[14]	Limited data (numeric and nominal).	FSDT + user profiling and sensing mechanism.	Bridge the gap between AI and human -like learning.
[16]	Unstructured, limited nominal data ("Book categories") and numeric data	Measuring quantitative ML explanations to cope with trustworthiness (LIME and COVAR).	Accuracy of 95.6% with LIME and 95.9% with COVAR.

[16]	Various scenarios	Incremental model to overcome the lack of data by defining trustworthy properties (trustee, prudential reasons, etc.).	Infer moral goals for end users.

B. Explainable machine learning

In this section, a new categorization of ML models is given based on the current research flows towards ML trustworthiness [18], [19], [20]. A further step has been taken to examine the user action after the prediction generation from ML models. Deep Neural Nets (DNNs) are particularly targeted by this approach, because with classification or clustering algorithms, there exists some techniques to ensure the same behavior of the trained models (e.g., Chi-square [21], features selection and cross validation [22], etc.). However, DNNs have complex structure (many hidden layers, parameters, weights, etc.), which makes the task of explaining predictions almost impossible. Technically, ML explanation is an additional layer between the user/expert and the learning process API that provides more insights about the predicted output. Local Interpretable Model-Agnostic Explanations (LIME) [23] is an explanation algorithm which covers more the interpretable side of any ML classifier. An intuition layer is presented in order to give a clear separation of the learning features and the remaining model by using distance function. This model has been refined to a selective method: Sparse Linear LIME (SP-LIME) to guarantee the model consistency while preserving a part of human logic.

DARPA program [18], [24] highlights a new learning process, which aims to simplify the ML models to increase users' satisfaction by preserving as much accuracy as possible. It consists of two additional layers: new simplified learning and explainable layer, see Figure 2.

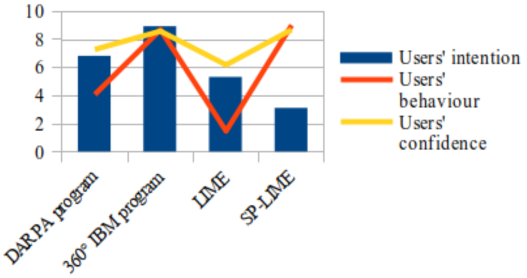

Figure 2. Explainable ML impact on user's reactions.

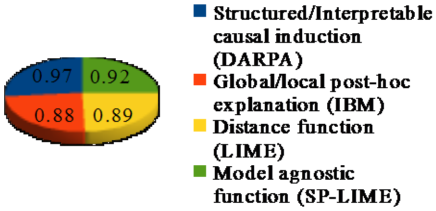

Figure 3. Explainable ML: techniques and accuracy.

IBM [25], [26] have published the 360 explainable AI which recognized users from different expectations and follow the domain expressiveness, a nurse for instance doesn't expect usually the same explanation from a surgical robot as with a neuroscientist. Figure 3 shows some techniques and their relative accuracy used in this kind of learning.

C. Explainable deep learning

As DNNs are getting more and more attention, deciphering their inner working mechanism has been subject of many studies [27], [28], [29], [30]. Unlike ML interpretability, explainable DNNs is much more challenging to be limited around clarifying the learning function itself [31]. However, as illustrated in Figure 4, more computational units have been included to support that.

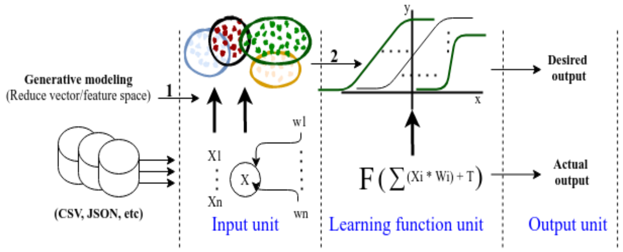

Figure 4. Generative modeling (1) and Post-hoc method (2) for DNNs' explanation.

Overall, there are two main approaches to proceed DNNs explanation:

- Generative modeling approach: which is depicted by (1) in Figure 4, it consists of inferring new correlations among input data, which are less complex [32]. The latter has the ability to reduce the samples space as well as the processing complexity [33] and to produce accurate predictions.
- Post-hoc methods require further processing than the first approach [34], it is about training the algorithm and try to improve the activation function based on previous inferred correlations as well as the primary (actual) output.

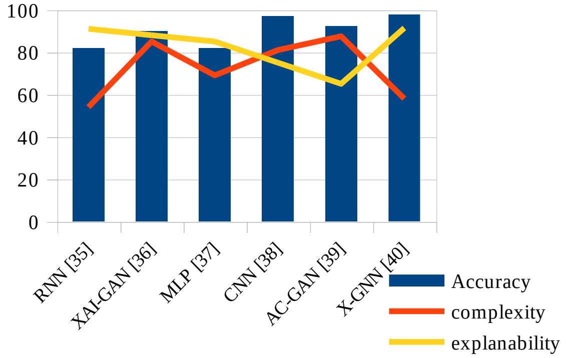

Figure. 5 Explanation, accuracy and complexity rates of Recurrent Neural Net (RNN), Generative Adversarial Net (GAN), Multi-Layer Perceptron (MLP), Convolutional Neural Net (CNN), and Graph Neural Network (GNN).

Since image processing has been dominating the field of deep learning the last decade [41], explainable Convolutional

Neural Networks (CNNs) have been widely investigated by preserving the back-propagation strategy [42], Figure 5 depicts some recent deep learning techniques, their explainable rates and their respective accuracies. What is noticeable is that when propagating (e.g., CNN, RNN) the model shows high accuracy thanks to the gradient optimization, but that increases the complexity because it implies an additional explanation layer due to vanishing gradients [43]. Preserving a good trade-off between the above illustrated evaluation metrics is still subject of research.

III. Evaluation

By the following, we want to highlight the advocated metrics addressed by our analysis through a comparative study (TABLE III). A box marked as ✓ means that the evaluation metric has been emphasized in the corresponding survey.

TABLE III
THE PROPOSED ANALYSIS COMPARED TO OTHERS BASED ON RELEVANT METRICS.

	Trust constructs			Structural units of the learning models			Users' perception		
Metrics/units	Complexity |	interactivity |	reliability	Input |	decision |	output	Intention |	Behavior |	confidence
Author(s)									
[44]	✓			✓	✓	✓			✓
[45]	✓	✓	✓		✓	✓	✓	✓	
[46]	✓	✓	✓	✓	✓	✓			✓
[47]		✓	✓	✓	✓	✓		✓	✓
Our analysis	✓	✓	✓	✓	✓	✓	✓	✓	✓

Within the ML life cycle [47], it is very common to approach the concept of trustworthiness up to the evaluation and deployment phases. The literature definitions try to associate attributes like reliable distribution of features and model robustness [48]; the latter attribute considers data specific features (e.g., overfitting, bias, etc.) as opposite to reliability which concerns the model working (e.g., features' selection, model optimizer, etc.). However, this association is done within a separate categorisation of the model units (Figure 4), which prevents for instance unwanted bias elimination [49]. Through our investigation, the quantification (Figures 2 and 5) as well as the combination of the learning units (Figure 4) enable a concrete sampling of trust constructs (e.g., confidence); therefore, these new trust features could be trained (based on initial observations (see "1" in Figure 4) and then passed to the approximator (learning function) in order to infer a prediction. The whole ML model can benefit from the new formalized metrics (i.e., unbiased learning, the new observations may prevent vanishing problem, etc.). Based on our analysis, we can provide the following answers to the research questions:

1). Trust can have many dimensions within a model life cycle (e.g., robustness against input changes, sensitivity of functions in decisional unit, etc.). As we discuss a user-centered approach (Figures 1 and 2), interpretable/explainable decisions play a key role on users' reactions, which together form a trustworthy formalism.

2). As the ML models' behaviour change (see B), inner configurations like features' combination, may have a strong impact on users' behaviour because some of those features reflect trust metrics that could change the whole model performance.

3). Explainable AI provides an abstract way to approach human understanding from the model's logic, it aims a generalizable learning by being independent from the input data. As opposite, interpretable methods are example-dependent, they apply specific attributes (e.g., stereotypes), it is usually referred to the observed behaviour as "trustee" and decisional bloc as "trustor".

4). The suitability of models for transparency has strong dependence on their traceability (i.e., execution trace, reasoning trace). As stated in B and C, classifiers and regression models are quite understood due to the unique learning function. However, multiple layer models (e.g., DL) require additional artefacts (e.g., generative modeling (Figure 4)) to cope with each layer specifications.

5). Current trends have been emphasized in the next section, where model-based explainable learning is increasingly popular. Research in this area is empowered by transferable learning [50], the latter consists of generalizing reusable computational fragments of a model as an inductive application.

IV. Discussion

In this section, we first try to justify the variations of the contribution works referring to transparency in ML. Then, we critically discuss some gaps of the pre-analysed interpretable and explainable models on ML. As it can be seen from Figure 6 [51], while the majority of works have targeted interpretable

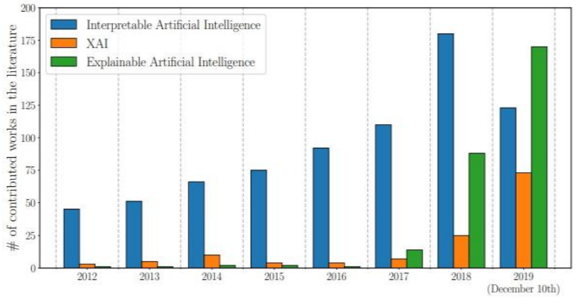

Figure 6. Variations of released research papers covering three main approaches of transparency in ML.

data driven techniques, this rate has shown a sudden decline in 2019, where latest contributions have been more driven by explainable AI; the latter has seen an increasing adoption rate during the last three years.

The contributions' rate adopting DARPA XAI technique has shown a fluctuating trend before 2017, as the project was not open source; however, it experienced a sudden increase during the last two years, but it remains lower than interpretable and explainable rates.

The previous arguments could be justified by the data driven available technologies and their high performance [52] on specific problems' evaluation (i.e., specificity, precision, etc.). As opposite to the model driven techniques (e.g., LIME, DARPA, Figure 4), where their application requires a domain expertise (i.e., local/global interpretation, post-hoc/generative modeling, etc.). However, the sudden increase of explainable AI by 2018 (Figure 6) follows the recent interest in inductive learning [53] and the emergence of abstraction methods (e.g., graph technologies [54], etc.).

A. Interpretable ML

- Interpretable ML approaches use data driven techniques (see A), the latter have improved ML accuracy and precision; however, they lack the users' behaviour and intentions that include experts and non-experts expertise toward trustworthiness.
- Models represent one component of the ML decision process, trust in ML cannot be only restricted to the model's interpretability based on specific attributes [13] or columns [15], it should cover the whole process according to the users' expectations. Thus, this abstract view may invoke a formalism in which a rigorous inference engine will cover the lack of expertise.
- In [13], combining fuzzy and ontological approach is an interesting way to justify learning metrics by satisfying the model hierarchy. However, the issue is that by having an initial model, users may not have complete view of its hierarchy which may generate a lack of understanding of these learning decisions, due to the absence of any mechanism which may infer missed concepts (features), inconsistency, mismatch, etc.

B. Explainable ML

The discussion here will focus on the behaviour of the ML systems shown in Figure 2. for LIME and SP-LIME projects, ML algorithms like decision-trees, linear, additive models can be traceable (path in trees, additive rules, etc.). However, what if a rule misses a critical feature as a user input mistake, how can this transparency be trustworthy.

Explainable 360 proposed by IBM has proven its effectiveness in several areas: medicine, finance, loans, etc. But, the explanation algorithm works mainly with predictive ML models while ML covers prescriptive, descriptive approaches [55].

Regarding DARPA's project, and based on [56] advocation, the critic is that while the predictions are well explained, it doesn't help to fix the issues occurring during the process. This argument is justified by an experience done with a number of patients when the ML model was collecting data from clinics instead of their medical records.

For deep learning models, post-hoc techniques turn these algorithms into interpretable models, but as it is covered, this is done approximately and those models lose their privacy [24] which is still critical for many systems. However, as shown by Figure 5, hybridizing the previous techniques with generative ones (e.g., XAI-GAN, XGNN, etc.) increases the models' performance despite their complexity.

V. Ethical and Legal Issues

In order to evaluate explainable AI policy, the first relevant question to ask is how far can we expend the learning systems transparency in accordance with liable and sensitive cases (e.g., in healthcare domain). These issues were discussed in [57], if a surgical robot bugs and kills someone or if a self-driving car hits a pedestrian, who should we blame? Even if a neural network usually provides accurate outcomes from patient records for instance, the lack of proof and verification techniques which are referred as 'Empathy' in [58] rises some ethical issues on how data has been trained and cleaned and which data had most influence on the prediction, for instance, etc.

VI. Conclusion and Future Research

This paper reviewed and analysed the recent studies on explainable and interpretable ML systems toward AI trustworthiness. ML and DL transparency in particular are increasingly emerging while ensuring a trade-off understandability/privacy, the latter is an important key of our discussion where in some cases a "Blackbox" model means a secured one. Through the analysis of several literature models, it has been noticed an exclusion of user's perception and admissibility metrics (i.e., intention, confidence, etc.) from ML and DL models' lifecycle. Therefore, it has been shown that a better understanding of the model components (input unit, decisional layers, function approximators, etc.) could reduce the gap between a model driven and a data driven explanation; which offers an easy integration of the discussed metrics into the same pipeline. In DNNs for instance, a batch of computation can be reused at the input space [40]; thus, the inclusion of the perceptual metrics could be achieved by employing an abstraction strategy (e.g., graph inference) as well as a way to infer missing concepts.

It is concluded that:

- Adding different explainable layers to learning models may be quite understandable for end-users (e.g., XDNN model [59]) but computationally expensive and not traceable.
- Modern explainable DL methods tries to stick with DL architecture and expand the explanation view to go beyond the learning function unit for better exploration of correlated inputs and desired outputs, embeddings techniques [60] showed promising results.
- Understanding users' psychology plays a key role toward trustworthy models; therefore, analysing their sentiments through DL may boost the understandabil-

ity of their inner working.

- Secure ML models do not mean trustworthy ones; however, in many cases, security means safety by which we entrust ML more in “critical” scenarios. There was a remarkable interest in data driven techniques [52] about designing security at earlier conceptual stages of ML.

This work could benefit from several potential directions:

- adopting logical reasoning into ML process may increase model certainty, the challenge is to figure out the right syllogism which mimics a learning theory; so that, it reduces the gap between example based and model generic explainability [61], [62].
- Considering AI policy [63] when formalizing the discussed metrics may help in certifying the consequences of a prediction regarding a certain behaviour or a perception. This could be useful when deciding to remove a disparate impact for instance without knowing the data bias, or even to justify a deletion of sparse data that could be sort of vanishing.

REFERENCES

[1] T. Liu, T. Qin, B. Shoa, W. Chen, and J. Bian, “Machine Learning: Research hotspots in the next ten years”. Microsoft Research Lab – Asia. Retrieved from https://www.microsoft.com/, 2018. Visited on 01/02/2020.

[2] N. Mahyar, V. D. Nguen, M. Chan, J. Zheng, and P. S. Dow, “The Civic Data Deluge: Understanding the Challenges of Analyzing Large-Scale Community Input”. DIS ’19, June 23–28, 2019, San Diego, CA, USA. Doi:http://doi.org/10.1145/3322276.3322354, 2019. Visited on 06/03/2021 10:16.

[3] R. Jacklin, N. Sevdalis, A. Darzi, and C. Vincent, “Mapping surgical practice decision making: an interview study to evaluate decisions in surgical care”. The American Journal of Surgery. 195(5):689–96, 2008.

[4] A. B. Farjadian, A. M. Annaswamy, and D. Woods, “Bumpless Reengagement Using Shared Control between Human Pilot and Adaptive Autopilot”. IFAC-PapersOnLine. Volume 50, Issue 1. pp. 5343-5348. doi https://doi.org/10.1016/j.ifacol.2017.08.925, 2017. Visited on 04/05/2019 12:50.

[5] D. Kakde, “Machine Learning Ensures Reliability for Space Missions: To the Moon and Beyond. DATAVERSITY Education”. Retrieved from https://www.dataversity.net/machine-learning-ensures-reliability-for space-missions-to the-moon-and-beyond/, 2019. Visited on 13/02/2020 20:31.

[6] M. Yin, J. M. Vaughan, and H. Wallach, “Undestanding the Effect of Accuracy on Trust in Machine Learning Models”. UK. ACM, New York, NY, USA, 12 pages. http://doi.org/10.1145/3290605.3300509, 2019. Visited on 14/09/2020 20:57.

[7] S. Liu, X. Wang, M. Liu, and J. Zhu, “Towards better analysis of machine learning models: A visual analytics perspective”. Visual Informatics1 48 56. doi: http://dx.doi.org/10.1016/j.visinf.2017.01.006 2, 2019.

[8] P. Ghoch, “Data Scientists and Machine Learning Algorithms for the Data- Driven World. DATAVERSITY”. Retrieved from htt ps://www.dataversity.net, 2018. Visited on 17/02/2020 02:36.

[9] F. Girardin and P. Fleurquin, “Designing for trust with machine learning”. The 2018 AAAI Spring Symposium Series, Association for the Advancement of Artificial Intelligence. Retrieved from www.aaai.org, 2018. Visited on 03/04/2019 08:45.

[10] J. Wang, X. Jing, Z. Yan, Y. Fu, W. Pedrycz, and L. T. Yang, “A Survey on Trust Evaluation Based on Machine Learning”. ACM Comput. Surv., Vol. 53, No. 5, Article 107, 2020. DOI:https://doi.org/10.1145/3408292

[11] D. Trček, “A formal apparatus for modeling trust in computing environments”. Mathematical and Computer Modelling 49 226–233. journal homepage: www.elsevier.com/locate/mcm, 2019. Visited on 15/05/2020 14:45.

[12] J. Zhang and X. Fang, “A computational trust fraimework for socal computing (a position paper for panel discussion on social computing foundations)”. 2010 IEEE Second International Conference on Social Computing, 2010.

[13] H. Fang, J. Zhang, and M. Sensoy, “A generalized stereotype learning approach and its instantiation in trust modeling. Volume 30, July–Au gust 2018, Pages 149-158. doi:https://doi.org/10.1016/j.elerap.2018.06.004, 2018. Visited on 21/03/2019 21:40.

[14] H. Fang, J. Zhang, and M. Sensoy, “A 2020 perspective on “A generalized stereotype learning approach and its instantiation in trust modeling. Electronic Commerce Research and Applications 40:100955, DOI:10.1016/j.elerap.2020.100955, 2020.

[15] F. Doshi-Velez and B. Kim, “Towards A Rigorous Science of Interpretable Machine Learning”. arXiv:1702.08608v2 [stat.ML] , 2017.

[16] P. Shmidt and F. Biessmann, “Quantifying Interpretability and trust in Machine Learning Systems". Association for the Advancement of Artificial Intelligence, 219. Retrieved from www. aaai.org, Visited on 04/08/2020 01:12

[17] A. Ferrario, M. Loi, and E. Viganò, “In AI We Trust Incrementally: a Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions”. Philosophy & Technology. Retrieved from https://doi.org/10.1007/s13347-019-00378-3, 2019. Visited on 17/02/2019 17:21.

[18] D. Gunning and D. W. Aha, "DARPA’s Explainable Artificial Intelligence Program”. Association for the Advancement of Artificial Intelligence, AI MAGAZINE. ISSN 0738- 4602, 2018.

[19] M. Harradon, J. Druce, and B. Ruttenberg, “Causal Learning and Explanation of Deep Neural Networks via Autoencoded Activations”. arXiv preprint. arXiv:1802.00541v1 [cs.AI]. Ithaca, NY: Cornell University Library, 2018.

[20] G. Klein, “Explaining Explanation, Part 3: The Causal Landscape”. IEEE Intelligent Systems 33(2): 83–88. doi.org/10. 1109/MIS.2018.022441353, 2018.

[21] X. Jin, A. Xu, R. Bie, and P. Guo, “Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles”. In: J. Li, Q. Yang, A. H. Tan. (eds) Data Mining for Biomedical Applications. Lecture Notes in Computer Science, vol 3916. Springer, Berlin, Heidelberg, 2006.

[22] R. Hussain and R. H. Ajaz, “Seed Classification using Machine Learning Techniques. Journal of Multidisciplinary Engineering Science and Technology (JMEST)”. Vol. 2 Issue 5, May – 2015. ISSN: 3159-0040.

[23] M. T. Ribeiro, S. Singh, and C. Guestren, “Why Should I Trust You? Explaining the Predictions of Any Classifier”. KDD 2016 San Francisco, CA, USA. doi: http://dx.doi.org/10.1145/2939672.2939778, 2016.

[24] M. Turek, “Explainable Artificial Intelligence (XAI). DEFENCE ADVANCED RESEARCH PROJECT AGENCY. Retrieved from https://www.darpa.mil/program/explainable-artificial-intelligence, 2019.Visited on 06/03/2020 20:11.

[25] A. Mojsilovic, “Introducing AI Explainability 360. Retrieved from https://www.ibm.com/blogs/research/2019/08/ai-explainablility-360/, 2019. Visited on 03/02/2019 01:12.

[26] H. Michael. “XRDS: Crossroads, The ACM Magazine for Students AI and Interpretation”, Volume 25 Issue 3, Spring 2019, Pages 16–19, 2019.

[27] R. K. Singh, R. Pandey, and R. N. Babu, “COVIDScreen: explainable deep learning framework for differential diagnosis of COVID-19 using chest X- rays. Neural Comput & Applic, 2021. https://doi-org.libaccess.hud.ac.uk/10.1007/s00521-020-05636-6. Visited on14/01/2021 20:41.

[28] G. Stephan, “A Path Toward Explainable AI and Autonomous Adaptive Intelligence: Deep Learning, Adaptive Resonance, and Models of Perception, Emotion, and Action”, 2020. Frontiers in Neurorobotics, Lausanne. DOI:10.3389/fnbot.2020.00036.

[29] E. Nigri, N. Ziviani, F. Cappabianco, A. Antunes, and A. Veloso, "Explainable Deep CNNs for MRI-Based Diagnosis of Alzheimer’s Disease," 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, United Kingdom, 2020, pp. 1-8, doi: 10.1109/IJCNN48605.2020.9206837.

[30] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller. “Causability and explainability of artificial intelligence in medicine”. https://doi.org/ 10.1002/widm.1312, 2020. Visited on 03/01/2020.

[31] A. SHRESTHA and A. A. MAHMOOD, “Review of Deep Learning Algorithms and Architectures. IEEE Access, PP(99):1-1. DOI: 10.1109/

ACCESS.2019.2912200, 2019.

[32] R. Ramapuram, M. Gregorova, and A. Kalousis. “Lifelong generative modeling. Neurocomputing”. Volume 404, pp 381-400. https://doi.org/10.1016/j.neucom.2020.02.115, 2019. Visited on 25/08/2020 23:53.

[33] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. Nature Machine Intelligence, 1 (5), pp.206-215, 2019.

[34] A. Heuillet, F. Couthouis, and N. Díaz-Rodríguez, “Explainability in Deep Reinforcement Learning”. Knowledge-Based Systems. Vol 214, 28 February 2021, 106685. https://doi.org/10.1016/j.knosys.2020.106685. Visited on 05/04/2021 12:35.

[35] B. J. Hou and Z. H. Zhou, “Learning With Interpretable Structure From Gated RNN”. IEEE Trans Neural Netw Learn Syst. 2020 Jul;31(7):2267-2279. doi: 10.1109/TNNLS.2020.2967051. Epub 2020 Feb 13. PMID: 32071002.

[36] V. Nagisetty, L. Graves, J. Scott, and V. Ganesh, “xAI-GAN: Enhancing Generative Adversarial Networks via Explainable AI Systems”. arXiv:2002.10438v2 [cs.LG] 26 Oct 2020.

[37] E. Zihni, V. I. Madai, M. Livne, I. Galinovic, A. A. Khalil, J. B. Fiebach et al. “Opening the black box of artificial intelligence for clinical decision support: A study predicting stroke outcome”. PLoS ONE 15(4): e0231166. https://doi.org/10.1371/journal.pone.0231166, 2020. Visited on 14/04/2021 16:45.

[38] I. Rio-Torto, K. Fernandes, and L. F. Teixeira, “Understanding the decisions of CNNs: An in-model approach”. Pattern Recognition Letters.Vol 133, pp 373-380. https://doi.org/10.1016/j.patrec.2020.04.004, 2020. Visited on 25/03/2020.

[39] S. Shao, P. Wang, and R. Yan, “Generative adversarial networks for data augmentation in machine fault diagnosis”. Computers in Industry, 106:85-93. DOI:10.1016/j.compind.2019.01.00, April 2019.

[40] H. Yuan, J. Tang, X. Hu, and S. Ji, “XGNN: Towards Model-Level Explanations of Graph Neural Networks”. arXiv:2006.02587v1 [cs.LG], 2020.

[41] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, “Information Fusion”, 42, pp_ 146-157, doi 10.1016/j.inffus.2017.10.006, 2018.

[42] T. Kipf and M. Welling, “Semi-Supervised Classification with Graph Convolutional Networks”. arXiv preprint arXiv:1609.02907, 2016.

[43] A. H. Ribeiro, K. Tiels, L. A. Aguirre, and T. B. Schön, “Beyond exploding and vanishing gradients: analysing RNN training using attractors and smoothness”. ArXiv: 1906.08482[cs.LG], 2019.

[44] X. Huang, D. Kroening,W. Ruan, J. Sharp, Y. Sun, and E. Thamo, “A Survey of Safety and Trustworthiness of Deep Neural Networks: Verification, Testing, Adversarial Attack and Defence, and Interpretability”. arXiv:1812.08342v5 [cs.LG], 31 May 2020.

[45] E. A. Toreini, K. Aitken, K. P. L. Coopamootoo, C. Elliott, V. G. Zelaya, and A. van Moorsel, “Technologies for Trustworthy Machine Learning: A Survey in a Socio-Technical Context”. arXiv:2007.08911v1 [cs.LG], 2020.

[46] D. V. Carvalho, E. M. Perelra, and J. S. Cardoso, “Machine Learning Interpretability: A Survey on Methods and Metrics”. *Electronics 8*(8), 832. https://doi.org/10.3390/electronics8080832, 2019. Visited on 08/05/2020 14:47.

[47] E. A. Toreini, K. Aitken, K. P. L. Coopamootoo, C. Elliott, V. G. Zelaya and A van Moorsel, “The relationship between trust in AI and trustworthy machine learning technologies,” arXiv:1912.00782, 2020^{a}.

[48] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks”. In 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 39–57, 2017.

[49] D. Wei, K. N. Ramamurthy, and F. P. Calmon, “Optimized score transformation for fair classification,” in Proceedings of the International Conference on Artificial Intelligence and Statistics, Palermo, Italy, 2020.

[50] F. Yang, W. Zhang, L. Tao, and J. Ma,“Transfer Learning Strategies for Deep Learning-based PHM Algorithms”. Appl. Sci. 2020, 10, 2361; doi:10.3390/app10072361, 2020.

[51] B. B. Arrietaa, N. D´ıaz-Rodr´ıguezb, J. D. Ser, A. Bennetot, S. Tabikg, A. Barbadoh et al. “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI”. TECNALIA. P. Tecnologico, Ed. 700. 48170 Derio (Bizkaia), Spain, 2020.

[52] A. Fariha, A. Tiwari, and Meliou, A. “Conformance Constraint Discovery: Measuring Trust in Data-Driven Systems”. arXiv:2003.01289v4 [cs.DB], 2021.

[53] Y. Yang and L. Song, “Learn to Explain Efficiently via Neural Logic Inductive Learning”. ArXiv:1910.02481 [cs.AI].

[54] M. Gaur, A. Desai, K. Faldu, and A. Sheth, ”Explainable AI Using Knwledge Graphs”. CoDS-COMAD Conference, 2020.

[55] D. Bachar, “Descriptive, Predictive and Presctive Analytics Explained. LOGILITY: Planing Optimized”. Retrieved from https://www.logility.com/blog/descriptive-predictive-and-prescriptive-analytics-explained/, 2020. Visited on 18/03/2020 12:14

[56] K. Sukel, “Artificial Intelligence ushers in the era of superhuman doctors”. New Scientist, 2017. Retrieved from https://www.newscientist.com/article/mg23531340-800-artificial intelligence-ushers-in-the-era-of-superhuman-doctors/. Visited on 18/03/2020 22:01.

[57] P. Paudyal, “Should AI explain itself? or should we design Explainable AI so that it doesn’t have to”. Retrieved from: https://towardsdatascience.com/should-ai-explain itself-or-should-we-design-explainable-ai-so-that-it-doesnt-have-to-90e75bb6089e, (2019, March 4). Visited on 06/02/2019.

[58] P. Ferris, “An introduction to explainable AI, and why we need it”. FreeCodeCamp.org. Retrieved from website: https://www.freecodecamp.org/news/an-introduction-to-explainable-ai-and why-we-need-it-a326417dd000/, 2018. Visited on 24/03/2019 14:47.

[59] P. Angelov and E. Soares, “Towards explainable deep neural networks (xDNN). Neural Networks”. Vol 130, pp 185–194. https://doi.org/10.1016/j.neunet.2020.07.010, 2020.

[60] T. W. Cenggoro, R.A. Wirastari, E. Rudianto, M. I. Mohadi, D. Ratj, and B. Pardamean, “Deep Learning as a Vector Embedding Model for Customer Churn”. Procedia Computer Science. Vol 179, pp 624 – 631, 2021. https://doi.org/10.1016/j.procs.2021.01.048. Visited on 14/04/2021 10:05.

[61] Z. H. Zhou, “Abductive learning: towards bridging machine learning and logical reasoning”. Sci China Inf Sci, 2019, 62(7):076101. https://doi.org/10.1007/s11432-018-9801-4. Visited on 24/12/2019 17:28.

[62] W. Z. Dai, Q. Xu, Y. YU, and Z. H. Zhou, “Bridging Machine Learning and Logical Reasoning by Abductive Learning”. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, 2019. Retrieved from https://papers.nips.cc/ paper/8548-bridging-machine learning-and-logical-reasoning-by-abductive-learning.pdf. Visited on 06/30/2020 12:56.

[63] O. Dowden, “New strategy to unleash the transformational power of Artificial Intelligence”. Retrieved from https://www.gov.uk/government/ organisations/office-for-artificial-intelligence, 12 March 2021. Visited on 25/04/2021 22:45.