Towards Human Understandable Explainable AI

Hani Hagras

School of Computer Science and Electronic Engineering, University of Essex, Wivenhoe Park, Colchester, CO43SQ, UK

Abstract

The recent increases in computing power coupled with rapid growth in the availability and quantity of data have resulted in a resurgence of interest in the theory and applications of Artificial intelligence (AI). However, the use of complex AI algorithms like Deep Learning, Random Forests, etc., could result in a lack of transparency to users which is termed black/opaque box models. Thus, For AI to be confidently rolled out by industries and governments, there is a need for greater transparency in explaining the AI decision making process to users to generate “White /Transparent Box” models which can also be termed Explainable AI (XAI). The paper reviews the need for XAI, the efforts to realise XAI and some areas which needs further exploration (like type-2 fuzzy logic systems) to realise XAI systems which could be fully understood and analysed by the lay user.

1. Introduction

Artificial Intelligence (AI) aims to make machines capable of performing tasks which require human intelligence. AI comprises all Machine Learning (ML) techniques besides other techniques such as search, symbolic reasoning and logical reasoning, statistical techniques and behaviour-based approaches. As technology, and, importantly, our understanding of how our minds and nature surrounding us work has progressed, our concept of what constitutes AI has changed.

With the huge increase in the amount of digital information being generated, stored, and made available for analysis, AI will have an important role to play. One key reason for building AI systems is not just to match human performance but in some cases exceed it. This is evident in situations where hundreds of input are contributing to a given decision where the human intuition will focus on a small set of inputs and small set of interactions due to the difficulty in figuring out the complex relationship between numerous inputs and their interaction. There are huge incentives to use AI for business needs including opportunities for cost reduction, risk management, enhancing decision-making, productivity improvements as well as developing new products and services. AI is a major disruptor and is anticipated to transform many industries where it is being rapidly adopted for a range of applications in various industries including mobile applications, security systems, speech recognition systems, financial related industries, Internet of Things, smart cities, automotive, biological sciences, pharmaceutics, etc.

AI is a technology revolution which the regulators and participants hope will be inclusive and benefit everyone, not just a select few. However, the use of complex AI algorithms like Deep Learning, Random Forests, Support Vector Machines (SVMs), etc., could result in a lack of transparency to create 'black/opaque box' models [1]. The lack of transparency issues are not specific to deep learning, or complex models**,** where other classifiers, such as kernel machines, linear or logistic regressions, or decision trees can also become very difficult to interpret for high dimensional inputs [2]. Such black/opaque box models cannot tell why a system made a decision, they just provide an answer which the user can take it or leave it [3].

According to the Financial Stability Board [4], in the financial sector, the widespread use of opaque models (like Deep Learning techniques) can lead to the lack of interpretability or 'auditability' which can contribute to macro-level risks [4]. As stressed by the financial stability board [4], it is important that progress in AI is accompanied with further progress in the interpretation of algorithms' outputs and decisions. This may be important condition not only for risk management but also for greater trust from the general public as well as regulators and supervisors in financial services [4].

According to the UK Parliament AI committee [5] "the development of intelligible AI systems is a fundamental necessity if AI is to become an integral and trusted tool in our society". "Whether this takes the form of technical transparency, explainability, or indeed both, will depend on the context and the stakes involved, but in most cases we believe ***explainability*** will be a more useful approach for the citizen and the consumer" [5]. They also mention, "***We believe it is not acceptable to deploy any artificial intelligence system which could have a substantial impact on an individual's life, unless it can generate a full and satisfactory explanation for the decisions it will take***" [5]. "In cases such as deep neural networks, where it is not yet possible to generate thorough explanations for the decisions that are made, ***this may mean delaying their deployment for particular uses until alternative solutions are found*** "[5].

Hence, there is a need to move towards Explainable AI (XAI) to enable the widespread of responsible and trusted AI to achieve the needed great positive impacts on communities and industries all over the World.

2. What is Explainable AI

The concept of explainability sits at the intersection of several areas of active research in AI, with a focus on the following areas [6]:

- **Transparency**: We have a right to have decisions affecting us explained to us in terms, format and language we can understand [7].
- **Causality**: If we can learn a model from data, can this model provide us not only correct inferences but also some explanation for the underlying phenomena?
- **Bias**: How can we ensure that the AI system hasn't learned a biased view of the world based on shortcomings of the training data or objective function?
- **Fairness**: If decisions are made based on an AI system, can we verify that they were made fairly?
- **Safety**: Can we gain confidence in the reliability of our AI system without an explanation of how it reaches conclusions?

An XAI or Transparent AI or Interpretable AI is an AI whose actions can be easily understood and analysed by humans. XAI can be used to implement a social right to explanation [8]. Hence, XAI is envisaged to provide the following benefits:

- Transparency and Compliance: It provides an auditable record including all factors and associations related with a given prediction. This enables a business to meet compliance requirements and eliminates concern that the organisation is hiding or doesn't know how a machine is affecting an outcome of a critical decision
- Ensures that there is an auditable and provable way to defend algorithmic decisions as being fair and ethical.

Transparency rarely comes for free; there are often tradeoffs between how accurate an AI is and how transparent it is, and these tradeoffs are expected to grow larger as AI systems increase in internal complexity. The technical challenge of explaining complex AI models decisions is sometimes known as the interpretability problem according to [9]. According to [10], XAI should aim to create a suite of machine learning techniques producing more explainable models, while maintaining a high level of learning performance (high accuracy). In addition, XAI should have the ability to explain their rationale, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future. These

XAI models can be combined with state-of-the-art human-computer interface techniques capable of translating models into understandable and useful explanation dialogues for the end user.

Producing formats which can only be understood and analysed by AI experts does not address the abovementioned issues as this will not allow the stake holder to test and augment the generated models with their experience. Hence, XAI should produce formats and outputs which can be easily understood and analysed by the Lay user/expert in a given field. This will allow domain experts to test the given the system and easily augment it with their expertise. This will allow the users and stake holders to understand the AI's cognition and allow them to determine when to trust or distrust the AI [10]. This will allow to satisfy the abovementioned points of transparency and causality and address the system bias, fairness and safety.

3. Previous and Current Work

XAI is one of DARPA (USA Defence Advanced Research Projects Agency) programs expected to enable “third-wave AI systems” [11], where machines understand the context and environment in which they operate, and over time build underlying explanatory models that allow them to characterize real world phenomena. According to [11], the XAI concept is to provide an explanation of individual decisions, enable understanding of overall strengths & weaknesses, convey an understanding of how the system will behave in the future and convey how to correct the system’s mistakes. Fig.1a shows a summary as provided by [11] showing the existing AI techniques performance vs explainability where it is shown that black box models like Deep Learning give best prediction accuracy vs Decision Trees which provide higher explainability contrasted by prediction accuracy.

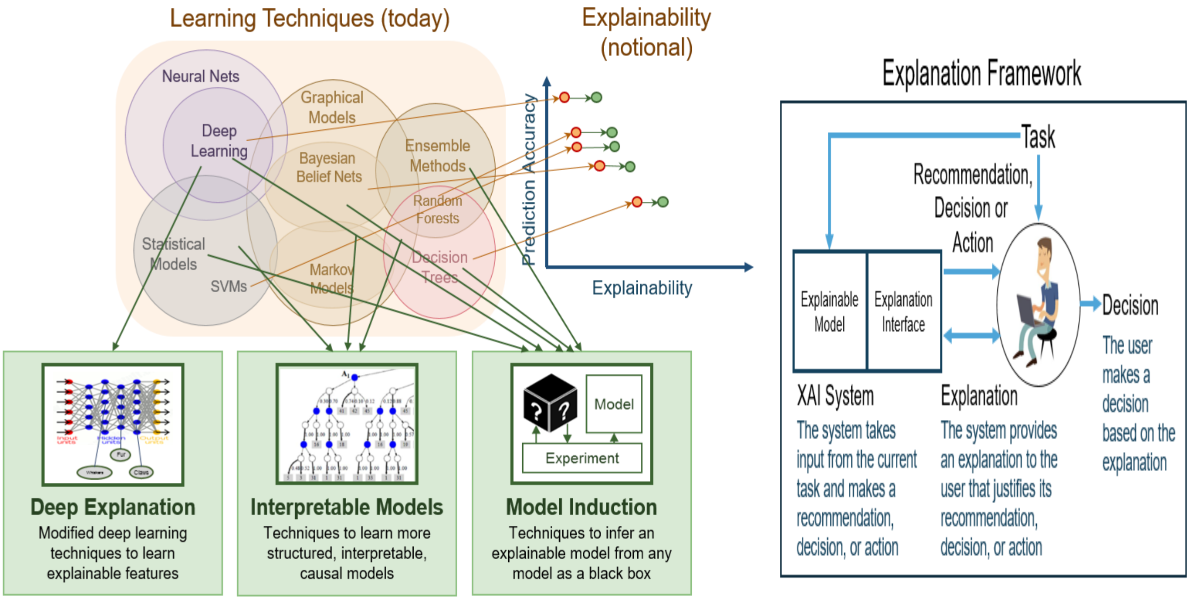

Learning Techniques (today)
Explainability (notional)
Neural Nets
Deep Learning
Graphical Models
Bayesian Belief Nets
Ensemble Methods
Random Forests
Statistical Models
SVMs
Markov Models
Decision Trees
Prediction Accuracy
Explainability
Deep Explanation
Modified deep learning techniques to learn explainable features
Interpretable Models
Techniques to learn more structured, interpretable, causal models
Model
Experiment
Model Induction
Techniques to infer an explainable model from any model as a black box
Explanation Framework
Task
Recommendation, Decision or Action
Explainable Model
Explanation Interface
Decision
The user makes a decision based on the explanation
XAI System
The system takes input from the current task and makes a recommendation, decision, or action
Explanation
The system provides an explanation to the user that justifies its recommendation, decision, or action

Fig1. a) Existing AI techniques- Performance vs Explainability [11]. b) The XAI explanation framework according to [11].

Decision Trees classify by step-wise assessment of a data point, one node at a time, starting at the root node and ending with a terminal node. At each node, only two possibilities are possible (left-right), hence there are some variable relationships that Decision Trees just can't learn. Although decision trees are usually considered easy to interpret, preparing decision trees, especially large ones with many branches, is complex and time-consuming. Large trees are not easily interpretable and pose presentation difficulties where it is quite difficult to analyse the common reasons and profiles pertaining to a decision where these entail the analysis of various routes and sub routes of the decision trees where the decision maker (and specifically the lay user) can be burdened with information slowing down decision-making capacity. This can be complicated further where there might be a possibility of duplication with the same sub-tree on different paths. Hence, although decision trees can be a good interpretable tool for problems with small number of features, they tend to be not easily read, explained and analysed (especially by the lay user) in problems with big number of features.

As shown in Fig 1a, in [11], they suggest various approaches to realise XAI, the first approach applies to Deep Learning and Neural Networks (which according to Fig. 1a and [11] have the highest predictive power) which is termed as deep explanation. This approach tries to modify the deep learning (or neural network) techniques to learn explainable structures. Some examples of such techniques can be found in [12] including, the Layer-wise Relevance Propagation (LRP) technique [13].

The second approach to XAI in Fig.1a is called interpretable models which are techniques to learn more structured and interpretable casual models which could apply to statistical models (e.g. logistic regression models, naïve bayes models, etc) , graphical models (such as Hidden Markov Models, etc) or Random Forests. However, like the deep explanation techniques, the output of these models could be analysed only by an expert in these techniques and not by a lay user.

The third XAI approach is what is termed model induction which could be applied to infer an interpretable model from any black box model [11]. According to [14], although it is often impossible for an explanation to be completely faithful unless it is the complete description of the model itself, for an explanation to be meaningful it must at least be locally faithful, i.e. it must correspond to how the model behaves in the vicinity of the instance being predicted. As mentioned in [14], local fidelity does not imply global fidelity: features that are globally important may not be important in the local context, and vice versa. While there are models that are inherently interpretable, an explainer (or model induction) should be able to explain any model, and thus be model-agnostic. An interpretable explanation need to use a representation that is understandable to humans, regardless of the actual features used by the model. In [14], a method was presented to explain a prediction output by sampling instances around *x'* (to create new point *z'*) by drawing nonzero elements of *X* uniformly at random. The method then aims to generate a model which is to be trained with *z and f(z)* [14]. In [14], they used sparse linear explanations, which lack the explanation of the interconnection between the various variables driving the given decision.

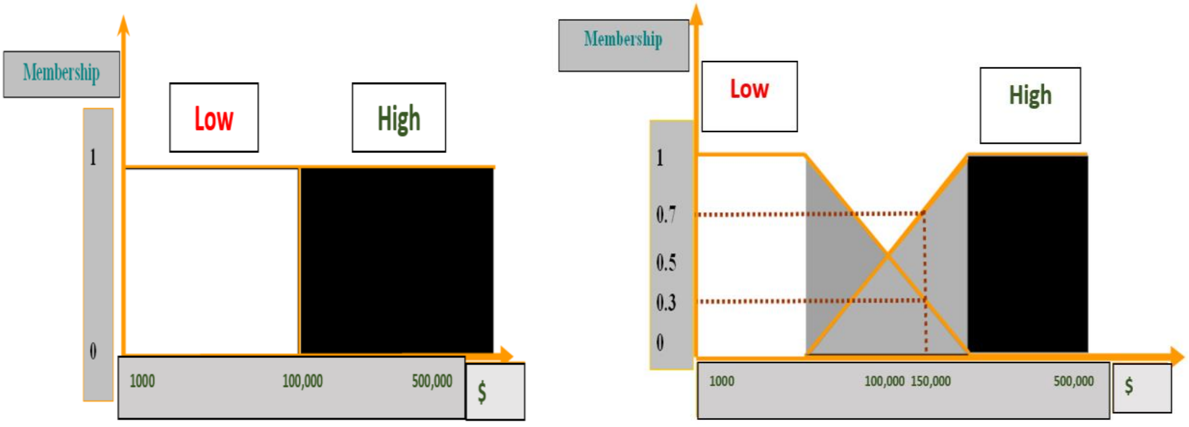
In [15], the same authors of [14] mentioned that explanations such as sparse linear models (used in [14]) can still exhibit high precision and low effort even for very complex models by providing explanations that are local in their scope. However, the coverage of such explanations are not explicit, which may lead to human error. As mentioned in [15], take the example of explaining a prediction of a complex model which predicts that the person described makes less than $50K. The linear explanation sheds some light into why, but it is not clear whether we can apply the insights from this explanation to other instances [15]. In other words, even if the explanation is faithful locally, it is not easy to know what that local region is [15]. Furthermore, it is not clear when the linear approximation is more or less faithful, even within the local region [15]. Hence in [15], they introduced Anchor Local Interpretable Model-Agnostic Explanations (aLIME) which is a system that explains individual predictions with crisp logic IF-Then rules in a model-agnostic manner. Such IF-Then rules are intuitive to humans, and usually require low effort to comprehend and apply [15]. In particular, an aLIME explanation (or an anchor) is a rule that sufficiently "anchors" a prediction – such that changes to the rest of the instance do not matter. For example, the anchor for this example might state that the model will almost always predict Salary < 50K if a person is not educated beyond high school, regardless of the other features. It was shown in [15] that the proposed approach outperform the linear based Model presented in [14]. However the IF-Then anchor model presented in [15], use crisp logic and thus will struggle with variables which do not have clear crisp boundaries, like income, age, etc. Also the approach in [15], will not be able to handle models generated from big number of inputs. Furthermore, explaining the prediction with just an anchor IF-Then rule does not give a full picture about the decision as for example in case of classification problems, there are always pros and cons which humans weigh in their minds and take a decision accordingly. Also, another major problem in an anchor approach, is the inability to understand the model behaviour in the neighbourhood of this instance and how the prediction can be changed if certain features could be changed, etc.

From the above discussion, it seems that offering the user with IF-Then rules which include linguistic labels appears to be an approach which can facilitate the explainability of a model output with the ability to explain and analyse the generated model as shown in Fig. 1b. One AI technique which employs IF-Then rules and linguistic labels is the Fuzzy Logic System (FLS). However, FLSs are not widely explored as an XAI technique and they donot appear in the analysis shown in Fig.1a. One reason might be is that FLSs are associated with control problems and they are not widely perceived as a ML tool as they need the help of other techniques to learn their own parameters from data. The following subsection will give an overview on FLSs and highlight their strengths and their misconceptions and present the type-2 FLSs as an important component to consider in the XAI developments.

4. Fuzzy Logic Systems and Human Understandable AI

Fuzzy Logic Systems (FLSs) attempt to mimic human thinking, although rather than trying to represent the brain’s architecture as you would with a neural network, the focus is on how humans think in an approximate rather than an precise way. A key facet of FLSs is in modelling and representing imprecise and uncertain linguistic concepts, creating a set of linguistic IF-Then rules to describe a given behaviour in human-readable form.

A good example would be the decision making process that a human goes through when they are driving a car. Rather than saying “if the distance to the car ahead is less than 2.5m **and** the road is 10% slippery **then** reduce car speed by 25%”, we would approximate the numerical elements with imprecise linguistic labels in the format of If the distance to the car ahead is ***low* and** the road is ***slightly slippery*** **Then *slow down.*** The numerical meanings of “low”, “close” and “slow down” will differ between drivers. Furthermore, if a driver was to be interviewed about the exact numerical values connected with these linguistic labels they would struggle to give a clear answer. Amazingly, humans are nevertheless able to communicate with these ill-defined and vague linguistic labels and do not query the exact values when they discuss them. In fact, these uncertain concepts allow humans to be able to perform very sophisticated tasks such as driving cars or underwriting financial applications.



(a) (b)

Fig.2. Representing the sets *Low* and *High* Annual Income using (a) Boolean sets. (b) Type-1 fuzzy sets.

Fuzzy Logic can model and represent imprecise and uncertain linguistic human concepts such as *Low*, *Medium*, *High*, etc. For example if a group of people were asked about the values they would associate with the linguistic concepts “*Low*” and “*High*” annual income and if Boolean logic was employed as shown in Fig. 2a then we would have to choose a threshold above which income values would be considered “*High*” and below which they would be considered “*Low*”. The first problem encountered is to identify a threshold that most people would agree on and this will be a problem as everyone has different idea what this linguistic label constitute. Even if an agreement was reached (say using a threshold of $100,000), will a value of $100,001 be considered “*High*” and will a value of $99,999 be considered “*Low*” income. It is clear that the hard boundary between the Boolean sets does not seem logical from a human point of view.

On the other hand the linguistic labels “*Low*” and “*High*” could be represented by employing the type-1 fuzzy sets. In this representation, it can be seen that no sharp boundaries exist between sets and that each value in the *x* axis can belong to more than one fuzzy set with different membership values. For example using Boolean logic, $150,000 used to belong only to the "*High*" set with a membership value of 1.0 in Fig. 2a. In Fig. 2b, using type-1 fuzzy logic, $150,000 belongs now to the "*Low*" and "*High*" sets but to different degrees where its membership value to "*Low*" is 0.3 and to "*High*" is 0.7. This can mean that if 10 people were asked if $150,000 is *Low* or *High* income, 7 out of 10 would say "*High*", (i.e. membership value of 7/10=0.7) and 3 out of 10 would say "*Low*", (i.e. membership value of 3/10=0.3). Hence, fuzzy sets provide a means of calculating intermediate values between absolute true and absolute false with resulting values ranging between 0.0 and 1.0, thus fuzzy logic allows the calculation of the shades of grey between true/false. In addition, the smooth transition between the fuzzy sets will give a good decision response when facing the noise and uncertainties. Furthermore, FLSs employ linguistic IF-THEN rules which enable to represent the information in a human readable form which could be easily read, interpreted and analysed by the lay user.

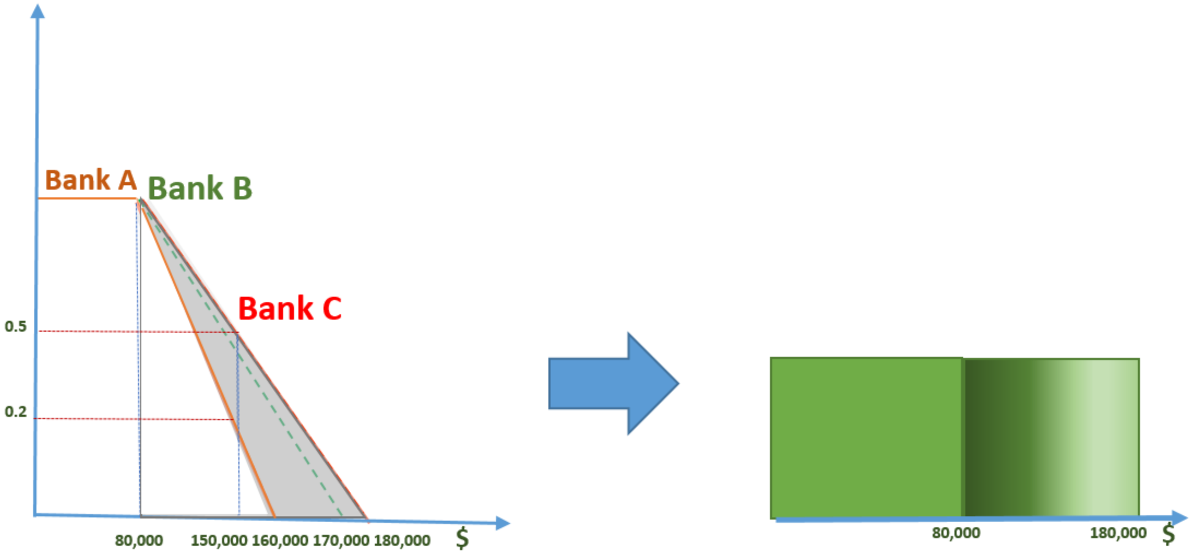
The type-1 fuzzy sets (shown in Fig.2b) are crisp and precise; hence they can handle only the slight uncertainties. However, different concepts mean different things to different people and in different circumstances and the memberships functions shown in Fig. 2b might vary in different countries, for different professions and across different underwriters in different banks. So assume as shown in Fig. 3a, we asked three financial experts in three different banks (Bank A, Bank B and Bank C) to cast their opinions about what are the suggested ranges for "*Low*" income. As can be seen in Fig. 3a, each expert might come with different type-1 fuzzy set to represent the "*Low*" linguistic label. Another way to represent linguistic labels is by employing type-2 fuzzy sets as shown in Fig 3a which embeds all the type-1 fuzzy sets for Bank A, Bank B and Bank C within the Footprint of Uncertainty (FoU) of the type-2 fuzzy set (shaded in grey in Fig. 3a). Hence, a type-2 fuzzy set is characterized by a fuzzy membership function, i.e. the membership value for each element of this set is a fuzzy set in [0,1], unlike a type-1 fuzzy set where the membership value is a crisp number in [0,1]. The membership functions of type-2 fuzzy sets are three dimensional and include a Footprint Of Uncertainty (FOU), this provide additional degrees of freedom that can make it possible to directly model and handle the uncertainties. In Fig. 3a, it can be seen that the $150,000 membership value to the set "*Low*" is no longer a crisp value of 0.3 as shown in Fig.2b, it is now a fuzzy function that takes values from 0.3 to 0.5 in the primary membership domain as shown in Fig.3a. More information about type-2 fuzzy sets and systems can be found in [16], [17].

One misconception about type-2 fuzzy sets is that they are difficult to understand by the lay person. However, this is not the case as if experts are questioned about how to quantify a linguistic label, they will be sure about a core value (which has a common consensus across all experts), however they will struggle to give exact points of the boundaries of this linguistic label and there will uncertainty about the end points of a given linguistic label. Hence, a simplified version of a type-2 fuzzy set can be shown in Fig. 3b where for the linguistic label "*Low*" income, there is a core value (shaded in solid green) of less than $80,000 which all experts agrees on and there is grey area (of shades of green) which goes between $80,000 to $180,000 of decreasing membership where there is uncertainty about the end points of the linguistic label where points beyond $180,000 are not recognised as "*Low*" income anymore.

Another misconception of FLSs in general is that they are control mechanisms. This is not true as the area of Fuzzy Rule-Based Systems (FRBSs) generated from data has been active for more than 25 years. However, this was hindered by the FLSs incapability to handle systems with big number of inputs due to the phenomena known as curse of dimensionality where the FLS can generate long rules and huge rule bases which turn them to black boxes which are not easy to understand or analyse. Furthermore, FRBSs werenot able to handle easily imbalanced and skewed data (such as those present in fraud, bank default data, etc). However, recent work such as [18], [19] was able to use evolutionary systems to generate FRBSs with short IF-Then rules and small number of rules in the rule base while maximizing the prediction accuracy. As this created sparse rule base not covering the whole search space, they presented a similarity technique to classify the incoming examples even if they do not match any fuzzy rule in the generated rule base. To do so, the similarity among the uncovered example and the rules was considered. They also presented multi-objective evolutionary optimization which was able to increase the interpretability (by reducing the length of each rule to include between 3 and 6 antecedents even if the system had thousands of inputs as well as having a small rule base) and maximize the accuracy of the FLS prediction. It was shown in [18], [19] that such highly interpretable systems outperform decision trees like C4.5 by a big margin in accuracy while being easy to understand and analyze than the decision trees counterparts.

What is most important is that unlike other white box techniques, the FRBS generates IF-Then rules using linguistic labels (which can better handle the uncertainty in information) where for example in a bank lending application a rule might be: IF Income is *High* and *Home Owner* and Time in Address is *High* Then *Good* Customer. Such rule can be read by any user or analyst. What is more important is that such rules get the data to speak the same language as humans. This allows humans to easily analyze and interpret the generated models and most importantly augment such rule bases with rules which capture their expertise and might cover gaps in the data (for example, human experience can augment such historically generated rules with the human expertise to cover situations which did not happen before). This allows the user to have full trust in the generated model and also cover all the XAI components mentioned in Section (2) related to **Transparency**, **Causality**, **Bias**, **Fairness** and **Safety**. Unlike the anchor rules mentioned in [15], humans do not make their decisions based on one single rule, they usually have Pros and Cons linguistic rules which humans balance and weigh in their mind and take a decision accordingly.

Hence, viewing Fig. 1a, it can be seen that type-2 FLS and FRBSs can be best in explainability while striking a good balance to prediction accuracy when compared to other black box techniques. Furthermore, the type-2 FLSs could be used to explain the decisions achieved from more complex black box modelling techniques. Hence, the type-2 FLS and FRBSs can offer a very good way forward to achieve XAI which can be understood, analysed and augmented by the lay user.



(a) (b)

Fig.3. a) A Type-2 fuzzy set embedding the type-1 fuzzy sets for the linguistic label "*Low Income*" from experts in three banks. B) A graphical simplification of the type-2 fuzzy set in Fig. 3a.

Acknowledgement

I would like to acknowledge Dr Robert Decaux from Logical Glue Ltd for his feedback on the paper.

References

[1] G. Nott, “Explainable Artificial Intelligence: Cracking open the black box of AI," Computerworld. Retrieved 2/11/2017.

[2] Z. Lipton, “The Mythos of Model Interpretability,” arXiv, June 2016

[3] A. Griffin, “Facebook's AI creating its own language is more normal than people think, researchers say," The Independent, 2017.

[4] “Artificial intelligence and machine learning in financial services," Available: http://www.fsb.org/wp-content/uploads/P011117.pdf

[5] “AI in the UK: ready, willing and able?,” UK Parliament (House of Lords) Artificial Intelligence Committee.
Available: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm

[6] C. Wierzynski, “The Challenges and Opportunities of Explainable AI,” Available: https://ai.intel.com/the-challenges-and-opportunities-of-explainable-ai/

[7] A. Weller, “Challenges for Transparency,” Arxiv, July 2017

[8] B. Goodman, S. Flaxman, “European Union regulations on algorithmic decision-making and a "right to explanation" “, AI magazine, Vol. 38, No.3, pp, 50-57, 2016.

[9] P. Voosen “How AI detectives are cracking open the black box of deep learning,” Science, 10.1126/science.aan7059, July 2017.

[10] A. Holzinger, M. Plass, K. Holzinger, G.Crisan, C. Pintea, V. Palade, Vasile. "A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop," ArXiv, 2017.

[11]D. Gunning, “Explainable Artificial Intelligence” http://www. darpa.mil/program/explainable-articial-intelligence, 2017.

[12] G. Montavon, W. Samek, K. Müller, “Methods for interpreting and understanding deep neural networks,” Digital Signal Processing, Vol.73, pp.1-15, 2018

[13] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, “On pixelwise explanations for non-linear classifier decisions by layer-wise relevance propagation,” PLoS ONE, Vol. 10, No. 7, 2015.

[14] M. Ribeiro, S. Singh, C. Guestrin, ”why should i trust you?”: Explaining the predictions of any classifier,” Proceedings of the 2016 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016

[15] M. Ribeiro, S. Singh, and C. Guestrin, “Nothing Else Matters: Model-Agnostic Explanations By Identifying Prediction Invariance," ArXiv e-prints, November 2016.

[16] H. Hagras, A Hierarchical Type-2 Fuzzy Logic Control Architecture for Autonomous Mobile Robots, IEEE Transactions on Fuzzy Syst. Vol. 12, No. 4, pp. 524-539, 2004.

[17] J. Mendel, “Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions”, 2nd Edition, Springer, January 2017

[18] J. Sanz, D. Bernardo, F. Herrera, H. Bustince, H. Hagras, “A Compact Evolutionary Interval-Valued Fuzzy Rule-Based Classification System for the Modeling and Prediction of Real-World Financial Applications with Imbalanced Data,” IEEE Transactions on Fuzzy Systems, Vol.23m No.4, pp.973- 990, August 2015.

[19] M. Antonelli, D. Bernardo, H. Hagras, F. Marcelloni, “Multi-Objective Evolutionary Optimization of Type-2 Fuzzy Rule-based Systems for Financial Data Classification,” IEEE Transactions on Fuzzy Systems, Vol. 25, No. 2, pp. 249-264, April 2017.