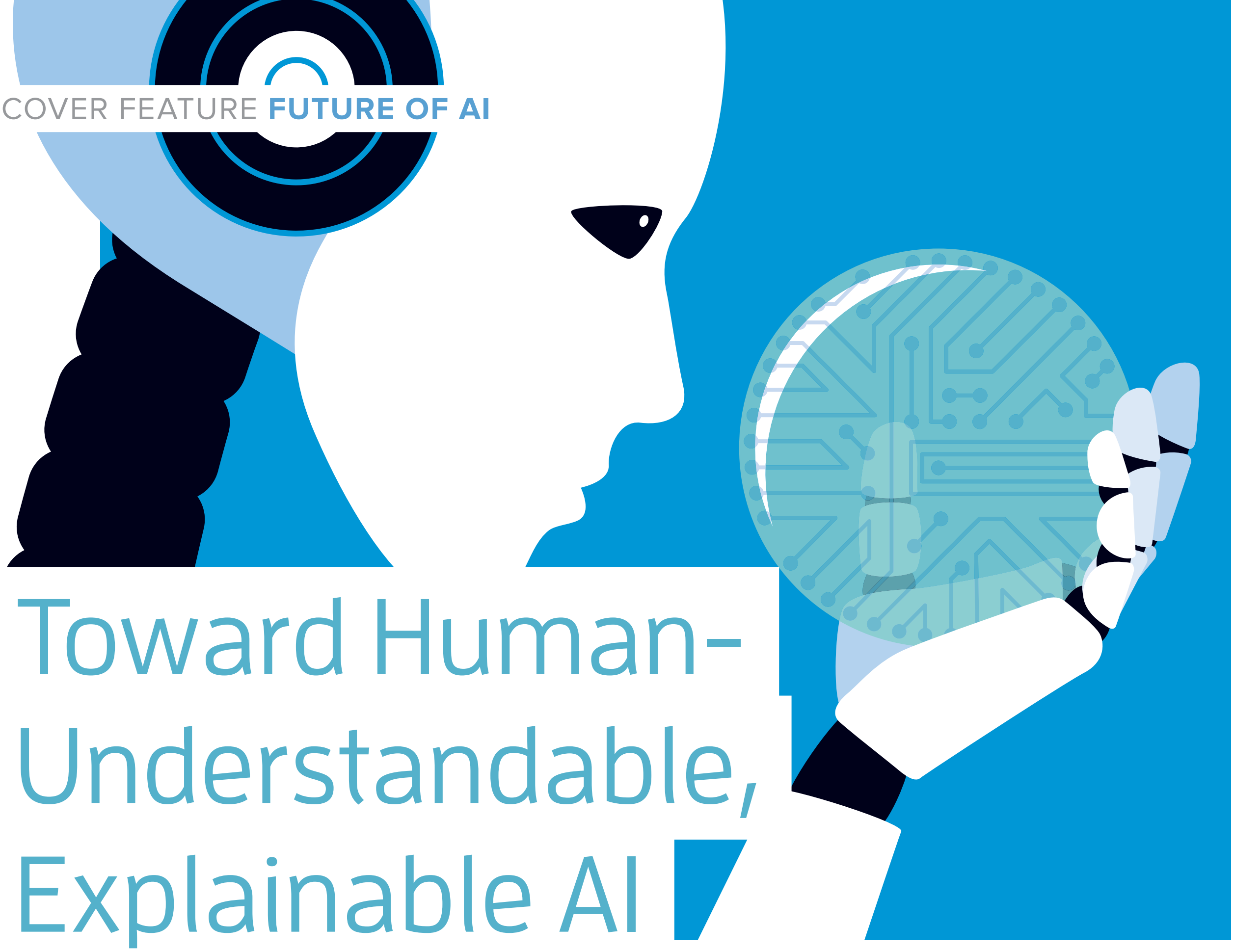

Toward Human-Understandable, Explainable AI

Hani Hagras, University of Essex

Recent increases in computing power, coupled with rapid growth in the availability and quantity of data have rekindled our interest in the theory and applications of artificial intelligence (AI). However, for AI to be confidently rolled out by industries and governments, users want greater transparency through explainable AI (XAI) systems. The author introduces XAI concepts, and gives an overview of areas in need of further exploration—such as type-2 fuzzy logic systems—to ensure such systems can be fully understood and analyzed by the lay user.

Artificial Intelligence (AI) is designed to make machines capable of performing tasks that usually require human intelligence. AI comprises all machine learning (ML) techniques as well as other techniques such as search, symbolic and logical reasoning, statistical techniques, and behavior-based approaches. As technology and, more importantly, our understanding of how our minds work and interact with all that surrounds us has progressed, our concept of AI has changed.

With the huge increase in the amount of digital information being generated, stored, and made available for analysis, AI has an increasingly 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

 0018-9162/18/$33.00 © 2018 IEEE

of inputs contribute to a given decision in which human intuition would focus on a small set of inputs and small set of interactions due to the difficulty of figuring out the complex relationships between numerous inputs and their interactions. There are huge incentives to use AI for business needs, including opportunities for cost reduction, risk management, enhanced decision-making, productivity improvements, as well as in the development of new products and services. AI is a major disruptor and is anticipated to transform those industries that are rapidly adopting it for a wide range of applications; these include mobile applications, security systems, speech recognition systems, financial industries, Internet of Things, smart cities, automotive technology, biological sciences, pharmaceutics, and more.

As the impetus for a technology revolution, AI is one for which the regulators and participants hope will be inclusive and benefit everyone, not just a select few. However, the use of complex AI algorithms such as deep learning, random forests, and support vector machines (SVMs), could result in a lack of transparency in order to create "black/opaque box" models.[1] These lack of transparency issues are not specific to deep learning, or complex models, there are other classifiers, such as kernel machines, linear or logistic regressions, or decision trees that 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 and the user can take it or leave it.[3]

According to the Financial Stability Board, which is an international agency that monitors global financial systems, the financial sector's widespread use of opaque models (like deep learning techniques) can lead to a lack of interpretability or "auditability," which could contribute to macro-level risks.[4] The Board issued a report in late 2017 that stressed how the progress in AI must be accompanied by further progress in the interpretation of algorithms' outputs and decisions.[4] This is an important condition, not only for risk management but also to establish greater trust from the general public as well as regulators and supervisors in financial services.[4]

According to a 2017 report from the AI Committee of the British Parliament,

> *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.... 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.... 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, it will be important to move toward "explainable AI" (XAI) to enable the widespread adoption of responsible and trusted AI and thereby achieve the significant positive impact on communities and industries all over the world.

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:[6]

- *Transparency*: We have a right to have decisions affecting us explained to us in terms, formats, and languages we can understand.[7]
- *Causality*: If we can learn a model from data, can this model provide us with not only correct inferences but also some explanation for the underlying phenomena?
- *Bias*: How can we ensure that the AI system has not 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 in which the actions can be easily understood and analyzed by humans. As presented in the work by Bryce Goodman and Seth Flaxman, XAI can be used to implement a social right to explanation.**[8]** Hence, XAI is anticipated to provide transparency and compliance, by providing 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 organization is hiding information or does not know how a machine is affecting the outcome of a critical decision, and to prove algorithmic decisions are 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 Paul Voosen.[9] According to Andreas Holzinger and his colleagues,[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 models 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 that can only be analyzed and understood by AI experts does not address the abovementioned issues as it does not allow the stakeholder to test and augment the generated models with their experience. Hence, XAI should produce formats and outputs which can be easily understood and analyzed 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 both users and stakeholders to understand the AI's cognition and empower them to determine when to trust or distrust the AI.[10] This establishes the ability to satisfy the abovementioned points of transparency and causality and address the system bias, fairness, and safety.

PREVIOUS AND CURRENT WORK

XAI is a DARPA program that is expected to enable "third-wave AI systems,"[11] in which 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 a 2016 DARPA report,[11] the XAI concept provides an explanation of individual decisions, enables understanding of overall strengths and weaknesses, and conveys an understanding of how the system will behave in the future and how to correct the system's mistakes. Figure 1a shows a summary as provided by DARPA, detailing the existing AI techniques' performance versus explainability in which it is shown that black box models like deep learning give the best prediction accuracy in comparison to decision trees, which provide higher explainability contrasted by prediction accuracy.

Decision trees classify by step-wise assessment of a data points, 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 or 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. In addition, it is quite difficult to analyze the common reasons and profiles pertaining to a decision when these entail the analysis of various routes and sub-routes of the decision trees and the decision maker (and specifically the lay user) can be burdened with information, thus slowing down decision-making capacity. This can be complicated further in cases in which there is 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 analyzed (especially by the lay user) in problems having large numbers of features.

As shown in Figure 1a, various approaches exist to realize XAI, the first approach applies to deep learning and neural networks (which are shown to have the highest predictive power in Figure 1a[11]), which is referred to 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 the work of Grégoire Montavon and his colleagues[12] including, the layer-wise relevance propagation (LRP) technique.[13]

The second approach to XAI in Figure 1a is *interpretable models*, which are techniques for learning more structured and interpretable casual models that could apply to statistical models (for example, logistic regression models, naïve Bayes models, and so on), graphical models (such as hidden Markov models, and so on) or random forests. However, like the deep explanation techniques, the output of these

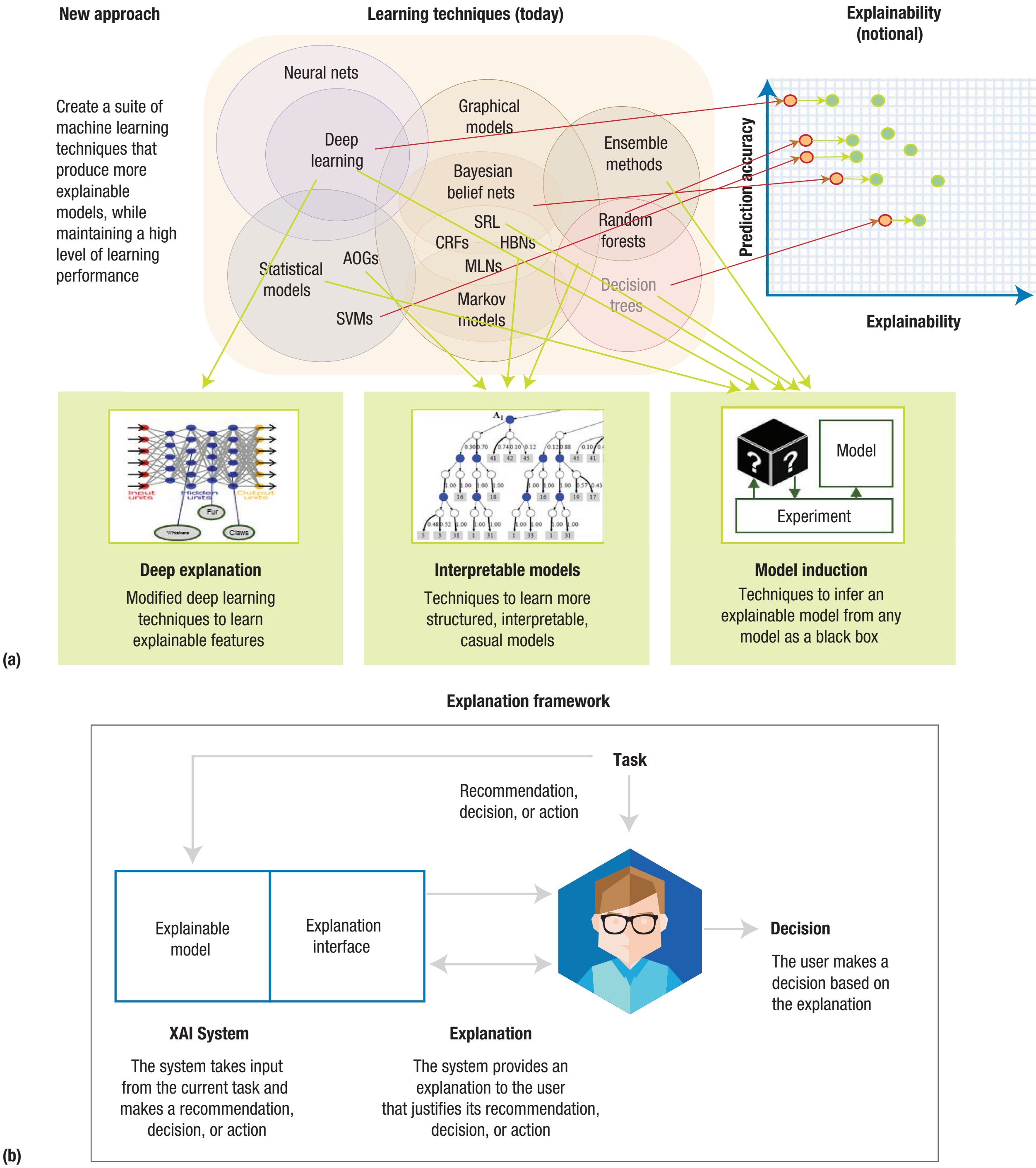

FIGURE 1. An overview of explainable models of AI (XAI). (a) Existing AI techniques, showing performance versus explainability. (b) The XAI explanation framework according to the 2016 DARPA report.[11] (Images courtesy of DARPA.)

models could be analyzed 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] The work of Marco Ribeiro and his colleagues has shown that 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—that is, it must correspond to how the model behaves in the vicinity of the instance beifng predicted.[14] In addition, 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 needs to use a representation that is understandable to humans, regardless of the actual features used by the model. Ribeiro and his colleagues presented a method 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 that is to be trained with z and $f(z)$.[14] They used sparse linear explanations, which lack the explanation of the interconnection between the various variables driving the given decision.

In another article,[15] Ribeiro and his colleagues describe how explanations such as sparse linear models 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 is not explicit, which can lead to human error. The authors also give an example[15] that explains a prediction of a complex model in which the person described makes less than \$50K; the linear explanation sheds some light on why, but it is not clear whether the insights from this explanation can be applied to other instances. 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. Hence, they introduced Anchor Local Interpretable Model-Agnostic Explanations (aLIME), which is a system that explains individual predictions with crisp *if-then* logic rules in a model-agnostic manner.[15]

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. Ribeiro and colleagues showed that the proposed approach outperforms[15] the linear based Model presented in their earlier work.[14] However, the *if-then* anchor model presented[15] uses crisp logic and thus it might be a struggle for handling variables that do not have crisp clear boundaries, such as income, age, and so on. Also, this approach will not be able to handle models generated from large numbers of inputs. Furthermore, explaining the prediction with just an anchor *if-then* rule does not give a full picture about the decision in the same way that, for example, classification problems will, there are always pros and cons that humans weigh in their minds to make the appropriate decision. Also, another major problem in an anchor approach, is the inability to understand the model behavior in the neighborhood of this instance and how the prediction can be changed if certain features change, and so on.

From the above discussion, it seems that offering users if-then rules that include linguistic labels appears to be an approach that can facilitate the explainability of a model output with the ability to explain and analyze the generated model as shown in Figure 1b. One AI technique that uses if-then rules and linguistic labels is the fuzzy logic system (FLS). However, FLSs are not widely explored as an XAI technique, and they do not appear in the analysis shown in Figure 1a. One reason for this might be 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.

FUZZY LOGIC SYSTEMS AND HUMAN-UNDERSTANDABLE AI

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 precise way. A key facet of FLSs is in modeling and representing imprecise and uncertain linguistic concepts, creating a set of linguistic if-then rules to describe a given behavior 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 2.5 m and the road is 10 percent slippery, then reduce car speed by 25 percent," we would approximate the numerical elements with imprecise linguistic labels in the format of *if* the distance to the car ahead is short and the road is slightly slippery, *then* slow down. The numerical meanings of "short," "slightly slippery," 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.

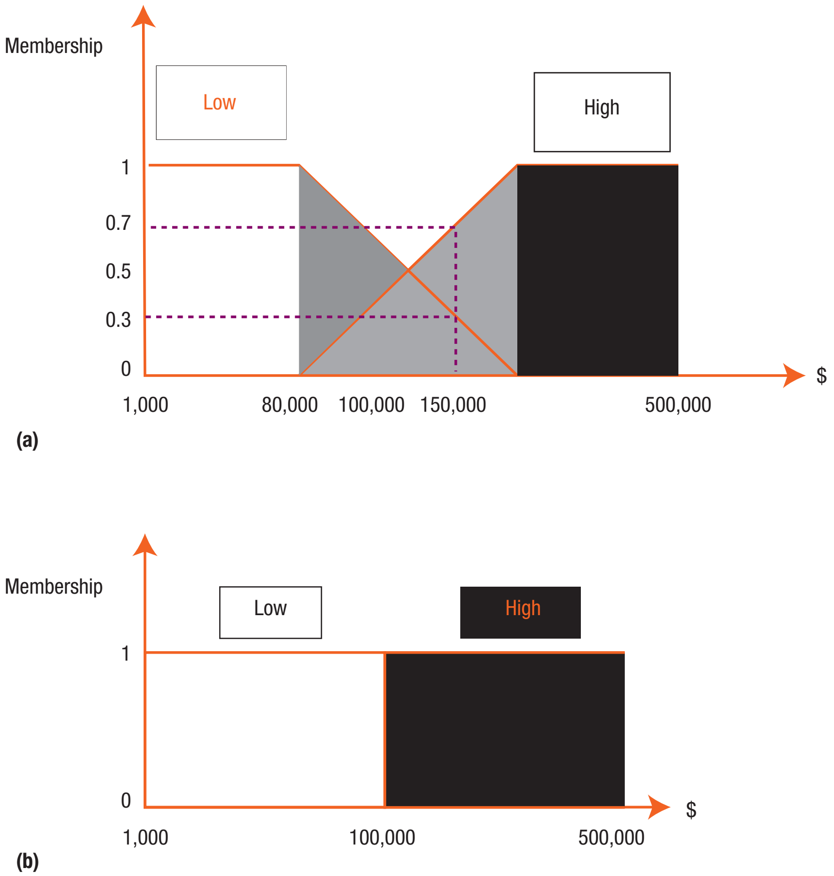

FIGURE 2. Representations of low and high annual income sets using (a) Boolean sets and (b) Type-1 fuzzy sets.

Fuzzy logic can model and represent imprecise and uncertain linguistic human concepts such as low, medium, and high. 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 Figure 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, which is a challenge as everyone has different ideas about what this linguistic label constitutes. Even if an agreement was reached (say using a threshold of $100,000), does this mean a value of $100,001 is considered high, but $99,999 is considered low? 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, linguistic labels *low* and *high* could be represented by employing the type-1 fuzzy sets. In this representation, no sharp boundaries exist between sets and each value on 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 Figure 2a. In Figure 2b, using type-1 fuzzy logic, $150,000 now 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

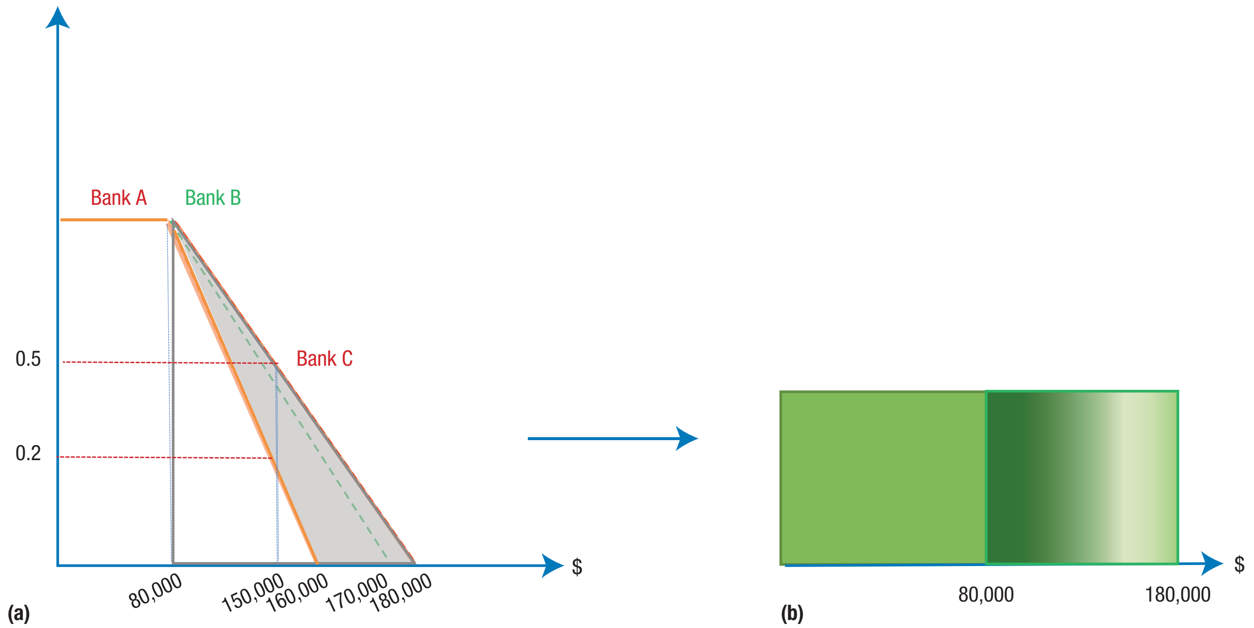

FIGURE 3. (a) A type-2 fuzzy set embedding the type-1 fuzzy sets for the linguistic label "low income" from experts hailing from three banks. (b) A graphic simplification of the type-2 fuzzy set shown in Figure 3a.

high (thus, a membership value of 7/10 is equal to 0.7), and 3 out of 10 would say *low* (again, membership value of 3/10 is equal to 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 and 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 that enable the information to be presented in a human readable form that could be easily read, interpreted, and analyzed by the lay user.

The type-1 fuzzy sets (shown in Figure 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 the memberships functions shown in Figure 2b might vary in different countries, for different professions and across different underwriters in different banks. So as shown in Figure 3a, we asked three financial experts from three different banks (Bank A, Bank B, and Bank C) to share their opinions about the ranges for *low* income. As can be seen in Figure 3a, each expert might come with a 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 Figure 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 gray in Figure 3a). Hence, a type-2 fuzzy set is characterized by a fuzzy membership function, specifically, 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 provides additional degrees of freedom that can make it possible to directly model and handle the uncertainties. In Figure 3a, it can be seen that the $150,000 membership value to the *low* set is no longer a crisp value of 0.3 as shown in Figure 2b, it is now a fuzzy function that takes values from 0.3 to 0.5 in the primary membership domain as shown in Figure 3a. More information about type-2 fuzzy sets and systems can be found in additional work.[16,17]

One misconception about type-2 fuzzy sets is that they are difficult to understand for the lay person. However, this is not the case 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 Figure 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 agree on and there is gray area (of shades of green) that goes between $80,000 and $180,000 of decreasing membership, and there is uncertainty about the end points of the linguistic label where points beyond $180,000 are not recognized as low income anymore.

Another misconception of FLSs in general is that they are control mechanisms. This is also 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 FLSs' incapability for handling systems with a 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 that are not easy to understand or analyze. Furthermore, FRBSs were not able to handle easily imbalanced and skewed data (such as those present in fraud, bank default data, and the like). However, recent work by Jose Sanz and his colleagues,[18] and Michela Antonelli and her colleagues,[19] used evolutionary systems to generate FRBSs with short if-then rules and a small number of rules in the rule base while maximizing the prediction accuracy. As this created a sparse rule base not covering the whole search space, they presented a similarity technique to classify the incoming examples, even if they did 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 that could 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. Previous work has demonstrated that such highly interpretable systems outperform decision trees like C4.5 by a big margin in terms of accuracy, while also being easier 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). So, for example, when a bank reviews a lending application, a rule might be: if income is *high* and *home owner* and *time in address* is *high*, then the application is deemed to be from a good customer. Such rules can be read by any user or analyst. More importantly, such rules get the data to speak the same language as humans. This allows us to easily analyze and interpret the generated models and, most importantly, to augment such rule bases with rules that 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 that have not happened before). This allows the user to have full trust in the generated model and also cover all the XAI components mentioned in the section related to transparency, causality, bias, fairness, and safety. Unlike the anchor rules mentioned in Ribeiro et al.,[15] humans do not make their decisions based on one single rule, they usually have pros and cons—linguistic rules that humans use to balance and

ABOUT THE AUTHORS

HANI HAGRAS is a professor, director of research, and director of the Computational Intelligence Centre at the School of Computer Science and Electronic Engineering, from the University of Essex, UK. His major research interests are in computational intelligence, notably type-2 fuzzy systems, fuzzy logic, neural networks, genetic algorithms, and evolutionary computation. Hagras received BSc and MSc degrees in electrical engineering from Alexandria University, Egypt, and a PhD degree in computer science from the University of Essex, UK. He is a Fellow of IEEE, and the Institution of Engineering and Technology (IET (IEE). He is a principal fellow of the UK Higher Education Academy (HEA). Contact him at hani@essex.ac.uk.

weigh in their mind and make a decision accordingly.

Hence, in viewing Figure 1a, it can be seen that type-2 FLS and FRBSs can be best in explainability, while striking a good balance with 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, analyzed and augmented by the lay user. C

REFERENCES

1. G. Nott, "Explainable Artificial Intelligence: Cracking Open the Black Box of AI," *Computerworld*, Retrieved 11 Feb. 2017; www.computerworld.com.au/article/617359/explainable-artificial-intelligence-cracking-open-black-box-ai.
2. Z. Lipton, "The Mythos of Model Interpretability," arXiv, June 2016; https://arxiv.org/abs/1606.03490.
3. A. Griffin, "Facebook's AI Creating Its Own Language Is More Normal than People Think, Researchers Say," *The Independent*, 3 Aug. 2017; www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-artificial-intelligence-ai-own-language-what-does-it-mean-fair-robots-chatbots-a7874576.html
4. "Artificial Intelligence and Machine Learning in Financial Services," report, Financial Stability Board, 1 Nov. 2017; www.fsb.org/wp-content/uploads/P011117.pdf.
5. "AI in the UK: Ready, Willing and Able?," report, UK Parliament (House of Lords) Artificial Intelligence Committee, 16 April 2017; https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm.
6. C. Wierzynski, "The Challenges and Opportunities of Explainable AI," Intel.com, 12 Jan. 2018; https://ai.intel.com/the-challenges-and-opportunities-of-explainable-ai/.
7. A. Weller, "Challenges for Transparency," Arxiv, 29 July 2017; https://arxiv.org/pdf/1708.01870.pdf.
8. B. Goodman and S. Flaxman, "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation,'" AI Magazine, vol. 38, no.3, 2016, pp., 50–57.
9. P. Voosen, "How AI Detectives Are Cracking Open the Black Box of Deep Learning," *Science*, 6 July 2017; doi: 10.1126/science.aan7059.
10. A. Holzinger et al., "A Glass-Box Interactive Machine Learning Approach for Solving NP-Hard Problems with the Human-in-the-Loop," ArXiv, 3 Aug. 2017; https://arxiv.org/abs/1708.01104.
11. D. Gunning, "Explainable Artificial Intelligence," report, Darpa, 11 Aug. 2016; www.darpa.mil/attachments/XAIIndustryDay_Final.pptx.
12. G. Montavon, W. Samek, and K. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, 2018, pp. 1–15.
13. S. Bach et al., "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, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *Proc. 2016 ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining* (KDD 16), 2016.
15. M. Ribeiro, S. Singh, and C. Guestrin, "Nothing Else Matters: Model-Agnostic Explanations by Identifying Prediction Invariance," ArXiv e-prints, November 2016; https://arxiv.org/abs/1611.05817.
16. H. Hagras, "A Hierarchical Type-2 Fuzzy Logic Control Architecture for Autonomous Mobile Robots," *IEEE Trans. Fuzzy Syst.*, vol. 12, no. 4, 2004, pp. 524-539.
17. J. Mendel, *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*, 2nd Edition, Springer, January 2017.
18. J. Sanz, et al., "A Compact Evolutionary Interval-Valued Fuzzy Rule-Based Classification System for the Modeling and Prediction of Real-World Financial Applications with Imbalanced Data," *IEEE Trans. Fuzzy Syst.*, vol. 23, no. 4, 2015, pp. 973- 990.
19. M. Antonelli, et al., "Multi-Objective Evolutionary Optimization of Type-2 Fuzzy Rule-based Systems for Financial Data Classification," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 2, 2017, pp. 249-264.