Paving the Way to Explainable Artificial Intelligence with Fuzzy Modeling

Tutorial

Corrado Mencar[1](✉) and José M. Alonso[2]

[1] Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro, Bari, Italy
corrado.mencar@uniba.it

[2] Centro Singular de Investigación en Tecnoloxías da Información (CiTIUS), Universidade de Santiago de Compostela, Santiago de Compostela, Spain
josemaria.alonso.moral@usc.es

Abstract. Explainable Artificial Intelligence (XAI) is a relatively new approach to AI with special emphasis to the ability of machines to give sound motivations about their decisions and behavior. Since XAI is human-centered, it has tight connections with Granular Computing (GrC) in general, and Fuzzy Modeling (FM) in particular. However, although FM has been originally conceived to provide easily understandable models to users, this property cannot be taken for grant but it requires careful design choices. Furthermore, full integration of FM into XAI requires further processing, such as Natural Language Generation (NLG), which is a matter of current research.

1 Introduction

Explainable Artificial Intelligence (XAI) is gaining consensus among researchers and engineers in Computer Science, as an alternative approach to current AI methods that show great learning capabilities but are relatively ineffective in explaining the reasons of the produced outputs in a human-intelligible way. Fuzzy Modeling has a huge potential for the development of advanced XAI systems, provided that some methodological requirements are fulfilled. The aim of this tutorial is to give a short overview of XAI and the way to reach it through Fuzzy Modeling in particular. After a brief introduction to XAI (Sect. 2), the role of Granular Computing is highlighted as the theoretical background that motivates the adoption of Fuzzy Modeling for XAI (Sect. 3). In particular, interpretability in Fuzzy Modeling is a key requirement for XAI, which is outlined in the subsequent Sect. 4. The next step toward XAI is the generation of natural language expressions to explain the decisions of a fuzzy (rule-based) model; NLG is briefly described in Sect. 5. Finally, some notes of possible future developments conclude this paper.

2 Towards Explainable Artificial Intelligence

In 2013, Eric Loomis was found driving a car that had been used in a crime. The judge sentenced him six-year of prison, which was determined in part by his

R. Fullér et al. (Eds.): WILF 2018, LNAI 11291, pp. 215–227, 2019.
https://doi.org/10.1007/978-3-030-12544-8_17

score on the COMPAS scale, an algorithmically determined assessment used to predict an individual's risk of recidivism. COMPAS is a proprietary algorithm, and its risk assessment procedure is opaque to the public. Loomis appealed against the sentence by objecting that the use of a predictive algorithm violated the principle of a due process but the Wisconsin Supreme Court ruled against Mr. Loomis because he would have gotten the same sentence based solely on the usual factors, including his crime and his criminal history [31][1].

Loomis' case is perhaps one of the first and most apparent examples of AI used to determine the course of a person's life. More and more cases accumulated in recent years, in very disparate situations, including autonomous vehicles, robot-assisted surgery, health-care, warfare, etc. AI is preponderantly entering our life and we must ask ourselves if we want this new presence and at which conditions.

The scientific community already recognized this new trend and began to react accordingly. In 2017, ACM issued a Statement on Algorithmic Transparency and Accountability which, by recognizing that computer algorithms have far-reaching impacts, their use may consciously or unconsciously result in harmful discrimination[2]. Accordingly, ACM recommends to use the same standards as institutions where humans have traditionally made decisions and outlines a set of principles, including the ability of explanation (a.k.a. *explainability*) which encourages to produce explanations regarding both the procedures followed by an algorithm and the specific decisions that are made.

From a political standpoint, the importance of data and their processing has recently been recognized and regulated. The General Data Protection Regulation (GDPR) is a EU regulation, emanated in 2016 and implemented in 2018, for the protection of natural persons with regard to the processing of personal data and on the free movement of such data[3]. GDPR is motivated, among other things by the right to obtain an explanation of the decision reached after any assessment provided by automatic procedures[4].

Explainable Artificial Intelligence (XAI) is a new approach to AI where the ability to explain the decisions provided by algorithms is the primary objective. The XAI program was firstly defined by the Defense Advanced Research Projects Agency (DARPA), with the objective of creating machine learning techniques that produce more explainable models, while maintaining a high level of prediction accuracy and «enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners»[5]. Figure 1 illustrates the differences between current AI (mainly based on Machine

[1] The full history has been reported by The New York Times, on May 2, 2017, p. A22. See https://nyti.ms/2qoe8FC.

[2] https://www.acm.org/binaries/content/assets/public-policy/2017_joint_statement_algorithms.pdf.

[3] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32016R0679.

[4] See note (71) in the preamble of GDPR. Actually, GDPR is quite timid in affirming the right of explanation [36], thence the need of more precise regulations on the subject in future.

[5] https://www.darpa.mil/program/explainable-artificial-intelligence.

Learning) and XAI according to DARPA: the learned function is replaced by an explainable model and an explainable interface for helping users understanding the results of a machine learning process.

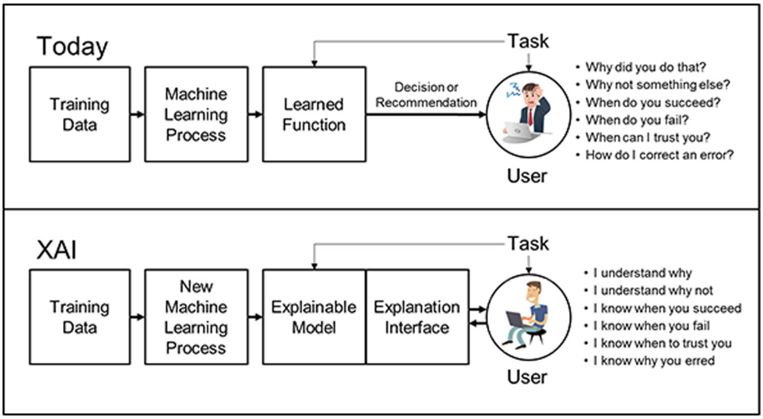


Fig. 1. XAI according to DARPA. Source: see footnote (5)

The importance of XAI is outstanding for several reasons, including: (i) the possibility of integrating machine and human knowledge in a simple way that is accessible by non-technical users; (ii) the possibility of interaction between users and machines in order to tackle complex problems; (iii) the ability of users to validate the functionality of an intelligent machine with respect to criteria of performance, ethics, safety, causality, etc.; (iv) the possibility of *trusting* machines for mission-critical applications [14].

XAI is growing widespread and reaching new frontiers on both scientific and technological sides. In this tutorial we will highlight the role of Granular Computing in general, and Fuzzy Modeling in particular, to the development of XAI.

3 Granular Computing

Granular Computing (GrC) is a computing paradigm where the object of processing is the *information granule*, i.e. a clump of objects kept together by some relations of indistinguishability, similarity, functionality or alike [44]. GrC is motivated by the need to approach AI through *human-centric information processing* [9], thence its central role in XAI.

GrC moves from some long-stated considerations concerning the apparent difficulty in developing common-sense reasoning in computers, while it seems so natural in human beings [28, Sect. 2.5]. These considerations led to the development of highly challenging branches of Informatics, such as Brain Informatics and

Cognitive Informatics, which aim at understanding the informational nature of human brain and mind by using the techniques provided by Informatics [37,45]. In particular, understanding the human brain and mind from the point of view of Informatics brought to a couple of fundamental assumptions: (i) brains and computers embody intelligence mostly for the same reasons; (ii) there exists a set of common principles that underlies both human intelligence and artificial intelligence [29]. Based on these assumptions, a theory of intelligence can be envisioned, which consists of multiple levels of explanations starting from the neural level up to the functional and conceptual level [38]. Deep neural networks are models of such a theory of intelligence which belong to the lowest neural and cortical levels. On the highest functional and conceptual levels, new forms of "human-inspired" computing models are needed; this need gave rise to GrC [38].

GrC is an "umbrella" paradigm that is declined in many forms according to the different branches of Artificial Intelligence. In particular, according to Zadeh, information granules are the results of granulation which, among organization and causation, are the three basic concepts of human cognition [44]. Specifically, granulation is the act of decomposing a whole into meaningful parts – like the decomposition of the image of a face into mouth, eyes, etc., or a satellite image into terrain, rivers, lakes, and so on.

Independent on the specific formal theories that can be developed under the paradigm of GrC, there are two common principles that are generally preserved: the *multilevel* and the *multiview* principles [39]. According to the multilevel principle, granulation yields a hierarchical granular structure, with levels in the hierarchy corresponding to different degrees of abstraction; on the other hand, each granular structure offers just a partial view of a phenomenon, therefore different granular structures (i.e. multiple views) may be used to provide a more complete understanding of the reality that is modeled. (A handy example is the scientific publishing model: title-abstract-content is a multilevel granular structure that is represented in a paper, and more papers are usually published on a subject to highlight the methodology, the application, the implementation, etc.).

Information granules at one level are treated as primitives for the higher level of a granular structure. Therefore, each information granule is informally defined as a collection of objects (i.e., information granules of the lower level) related together by some relation that makes objects indistinguishable at the higher level. Similarity, spatial proximity, functionality are examples of such relations.

Many concepts in the human mind are formed through an act of *perception*, i.e. the organization, identification and interpretation of a sensation in order to form a mental representation [32, Chap. 4]. Since what is perceived belongs to a continuous Reality and concepts are formed through perceptions, it is straightforward to assume that such concepts reflect the continuity of perceptions. Information granules are used to represent and process concepts as conceived by human minds, therefore information granules should be defined in order to preserve the continuity of perception-based concepts. Fuzzy Set

Theory (FST) offers a suitable mathematical underpinning to define this kind of information granules [43]. In other words, «fuzziness of information granules is a direct consequence of fuzziness of the concepts of indistinguishability, similarity, proximity and functionality» [40].

Very often, perception-based concepts are designated by labels forming our Natural Language [42]. Therefore, FST can be used for *Computing With Words* [41]: propositions in natural language are translated into fuzzy constraints on the involved variables; inference is carried out through the machinery offered by FST; the results of inference are eventually expressed in natural language. FST is a promising approach for defining the theoretical background to represent perception-based information granules, which are designated by linguistic terms drawn from natural language. Thus, FST is a natural candidate for designing models in XAI. In the next Section, we'll look at the opportunities and challenges deriving from the use of FST in XAI.

4 Interpretability in Fuzzy Modeling

Fuzzy Modeling (FM) is a methodology oriented toward the design of explanatory and predictive models using FST. FM is long-standing, with pioneering works dated in the seventies. The original intent of FM was to develop knowledge-based models capable of both representing highly non-linear relations between inputs and outputs, and at the same time offering an intelligible view of such relations through the use of a simplified natural language [22]. This was accomplished by "fuzzy rules"; nowadays, fuzzy rule-based models are common practice in FM.

In the eighties FST met Machine Learning [33,34], and since then several methods for automatically deriving fuzzy rule-based models from data arose. As a result, such fuzzy models were mainly designed for accuracy, while the original intent of FST to represent perception-based knowledge became of secondary relevance. But fuzzy rule-based models that are not *interpretable* are akin to black-box models, like neural networks, for which an armamentarium of powerful learning techniques already exist and are continuously refined. Interpretability is a property of fuzzy rule-based models which can be roughly defined as the capability of reading and understanding the knowledge-base of a (fuzzy) model. Interpretability is not given from grant by the mere use of FST but it requires a methodology that is still in development.

The definition of interpretability cannot be formulated in strict mathematical sense because it involves the human factor which is hard, if not impossible, to formalize. However, the basic principle underlying interpretability can be found in Michalski's Comprehensibility Postulate, which parallels the results of a learning algorithm with the description that a human expert might produce by observing the same entities [27]. Roughly speaking, the perception-based concepts acquired by a human should be *co-intensive* with the information granules that are automatically generated by a learning algorithm, provided that the same objects are observed [23]. In particular, since we use symbolic terms drawn from natural language to communicate knowledge, then the implicit semantics that a term conveys when used in a context should overlap with the explicit semantics a term is given by interpreting it with a fuzzy set (see Fig. 2).

It is possible to illustrate the concept of co-intension with an explanatory example involving two actors, Alice and Bob [24], where Alice is a scholar communicating some piece of information to Bob by adopting an appropriate language that is capable to represent her own knowledge. In order be understood by Bob, Alice chooses linguistic terms whose meanings are supposed to be shared by Bob. (It is not necessary that Alice's and Bob's meanings are exactly the same, but they should be overlapping enough in order to understand each other.) This is possible if Alice and Bob share similar environment, language, culture, experiences, etc. Therefore, co-intension can be achieved if Alice and Bob share similar conceptualizations of the piece of reality they are talking about. This illustrative scenario is very common among humans as it enables communication of information and knowledge. The comprehensibility Postulate tries to extend this principle to the communication of knowledge acquired by machines to humans.

Interpretability calls for both semantic and structural requirements, whereas the semantic facet is related to the co-intension of information granules with perception-based concepts, and the structural facet is needed to cope with the limited capabilities of the human brain in processing information [7]. In order to achieve an effective definition of interpretability, a collection of interpretability constraints and criteria can be adopted. This collection is not standardized, because different constraints can be selected according to the needs of the designer. As a consequence, there is not a unique computational definition of interpretability. It is common to organize interpretability constraints according to the level of modeling. Therefore, there are interpretability constraints for fuzzy sets, for linguistic variables, for multi-dimensional information granules, for fuzzy rules and for entire fuzzy models [25]. Assessment of interpretability is aimed at formalizing measures that quantify the degree of fulfillment of interpretability constraints by any model component. Also in the case of assessment, both structural and semantic measures are used and eventually aggregated to define a global evaluation of interpretability [16]. As an alternative approach, interview-based experiments can be used to evaluate the interpretability of a fuzzy rule-based model in a holistic way [5].

Designing interpretable fuzzy models requires some additional steps with respect to the usual modeling stages. In particular, the source of knowledge may be twofold: the available data and the expert's knowledge. The way of considering these two sources is critical for an effective model. An iterative approach is recommended to integrate induced knowledge with expert rules in order to drive the design toward a model that is balanced in terms of predictive and explanatory capabilities [6]. Also, several modeling approaches are available, which may favor interpretability over accuracy or the converse; other approaches recognize that interpretability and accuracy are conflicting objectives and adopt multi-objective techniques to achieve a Pareto front of solutions [12]. Alternatively,

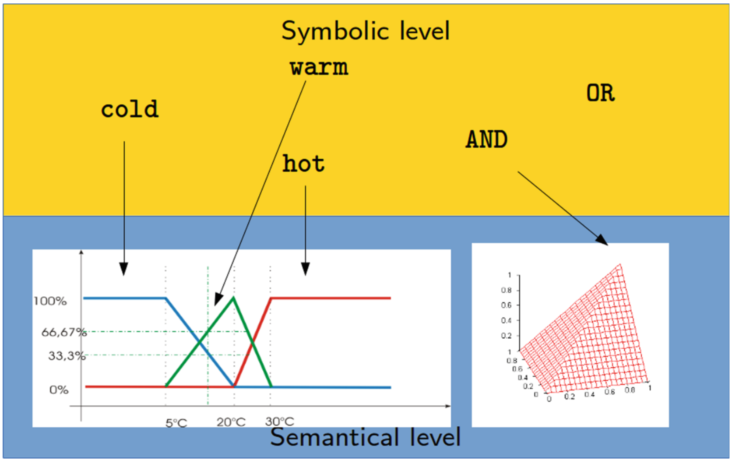


Fig. 2. Interpretation of symbols with fuzzy sets. Interpretability is assured only if the implicit semantics conveyed by each linguistic term is co-intensive with the explicit semantics determined by its interpretation.

ad-hoc algorithms may be used to incorporating interpretability constraints within the algorithms that induce fuzzy rules from data [13].

Most methods for modeling interpretable fuzzy systems adopt type-1 fuzzy sets, i.e. fuzzy sets that map objects of a universe of discourse into a scalar degree of membership. There is, however, a large corpus of literature concerning the use of type-2 fuzzy sets in fuzzy modeling. Type-2 fuzzy sets map elements of a universe of discourse into a type-1 fuzzy set defined on the domain of membership degrees. Type-2 fuzzy sets are justified by the assumption that «words mean different things to different people», therefore the uncertainty, related to the membership degree an object has to a set modeling a word, can only be represented by another level of uncertainty, thus giving rise to type-2 fuzzy sets [26]. Type-2 fuzzy sets gained attention in the last 15 years, not without complicacies and misconceptions [20]. For example, set operations on type-2 fuzzy sets can be defined in different ways, leading to very different theories [11]. Also, type-2 fuzzy sets may have different interpretations (e.g. in terms of intuitionistic or bipolar information). Therefore, type-2 fuzzy sets have a potential usefulness in modeling the meaning of words, but their manipulation and interpretation requires a full understanding of the subject of modeling. The authors' position is to favor type-1 fuzzy sets to model the knowledge base of a specific agent, while type-2 fuzzy sets are more suitable to model a kind of "social knowledge" that is shared among different agents. This is, however, matter of future research.

There are not many software tools to support designers in developing interpretable fuzzy models [1]. FisPro[6] is an open-source software that facilitates

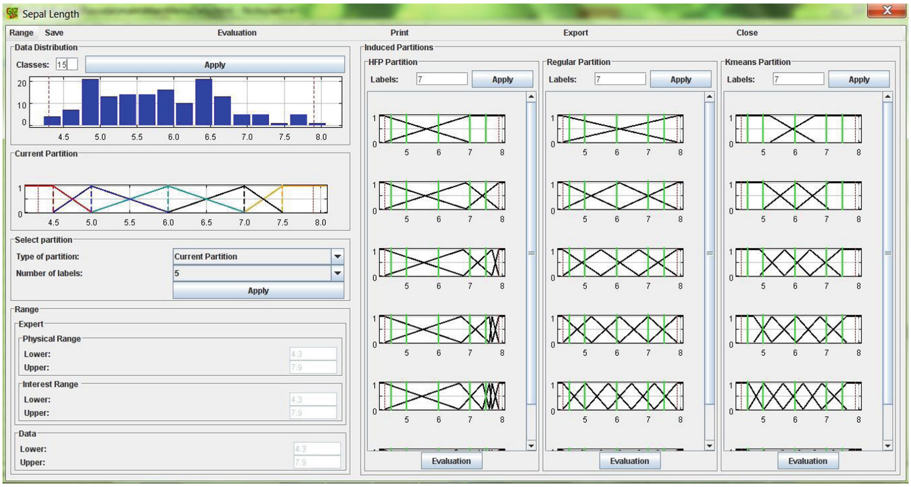
[6] https://www7.inra.fr/mia/M/fispro/fispro2013_en.html.

interpretability in all fuzzy modeling steps [19]. GUAJE[7] (Generating Understandable and Accurate fuzzy models in a Java Environment) is another open-source software with the aim of supporting the design of interpretable fuzzy rule-based systems by means of combining several preexisting software tools [2]. It is a portable graphical tool designed in order to facilitate knowledge extraction and representation for fuzzy rule-based systems, paying special attention to interpretability issues (see Fig. 3). GUAJE lets the user define expert variables and rules, but also provides supervised and automatic learning capabilities. Both types of knowledge, expert and induced, are integrated under the expert supervision for ensuring interpretability and consistency of the knowledge base along the whole process. The tool is an implementation of the HILK++ methodology for interpretable fuzzy modeling [4].

5 Fuzzy Modeling for XAI: Current Developments

Interpretability in fuzzy modeling is a requirement that leads to the development of methods and techniques to generate fuzzy models—mostly fuzzy rule-based systems—whose knowledge bases can be read and understood by users. In order to develop XAI, a step forward must be done, since in this case the new requirement is to explain the decision provided by a system. An interpretable fuzzy system gives the necessary information, but the explanation of a decision needs further processing.

XAI is a flourishing research direction is Artificial Intelligence, particularly in Machine Learning [10,18]; in Fuzzy Logic, research is gradually including the



Classes: 15

Type of partition: Current Partition

Number of labels: 5

Lower: 4.3

Upper: 7.9

Lower: 4.3

Upper: 7.9

Lower: 4.3

Upper: 7.9

Labels: 7

Labels: 7

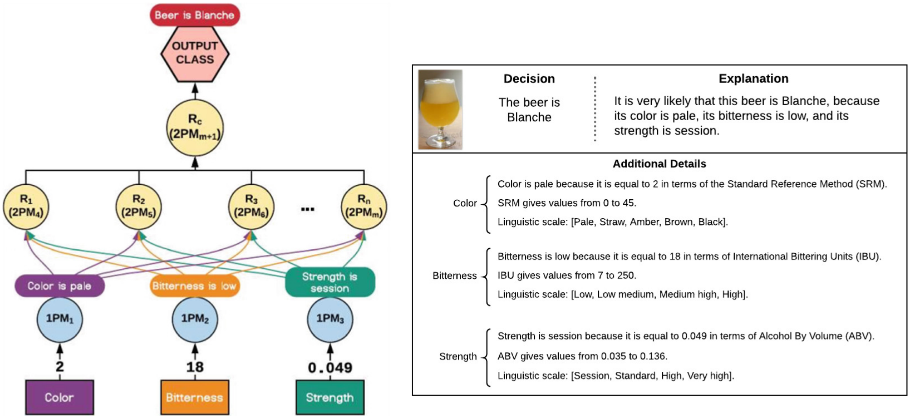
Labels: 7

Fig. 3. A screen-shot of GUAJE.

[7] https://sourceforge.net/projects/guajefuzzy/.

results in interpretability to develop explainable models based on fuzzy models [15]. A promising methodology that drives interpretable fuzzy modeling toward XAI is Natural Language Generation (NLG). NLG enables the generation of text from other data sources and finds application in state-of-art systems such as speech recognition, machine translation and conversational systems among others [17]. A specific branch of NLG is the so-called "data-to-text" (D2T-NLG), whereas linguistic descriptions are automatically generated from a complex of data. A particular approach for D2T-NLG is based on Linguistic Description of Complex Phenomena (LDCP), a method for NLG that produces a Granular Linguistic Model of a Phenomenon (GLMP), i.e. a network of processing units called "perception mappings", each of them representing a computational perception or an aggregation thereof [35]. A computational perception is a unit of meaning for the phenomenon under analysis and is identified by a set of linguistic expressions and their corresponding validity values given a situation (e.g. an input sample). Perception mappings aggregate computational perceptions by means of aggregation functions, which could be implemented in form of fuzzy rules, and generate appropriate text by an algorithm. The output of a GLMP is a linguistic description that explains a possibly complex situation, thanks to use of one or more underlying interpretable fuzzy models that are distributed among the perception mappings [3]. Figure 4 illustrates an example of GLMP used for generating an explanation of the inference carried out by a fuzzy rule-based classifier and the corresponding explanation for a given input sample.

A challenge in LDCP is to explain a phenomenon involving correlated data, whereas this relation has been learned by some inductive algorithm. A typical example is given by a Machine Learning algorithm that is used for learning a classification function: this algorithm could be highly accurate but it may hardly explain why a class label has been assigned to a given input. A possible approach is to use the classification algorithm as an oracle and a collection of interpretable



Decision

The beer is Blanche

Explanation

It is very likely that this beer is Blanche, because its color is pale, its bitterness is low, and its strength is session.

Additional Details

Color
- Color is pale because it is equal to 2 in terms of the Standard Reference Method (SRM).
- SRM gives values from 0 to 45.
- Linguistic scale: [Pale, Straw, Amber, Brown, Black].

Bitterness
- Bitterness is low because it is equal to 18 in terms of International Bittering Units (IBU).
- IBU gives values from 7 to 250.
- Linguistic scale: [Low, Low medium, Medium high, High].

Strength
- Strength is session because it is equal to 0.049 in terms of Alcohol By Volume (ABV).
- ABV gives values from 0.035 to 0.136.
- Linguistic scale: [Session, Standard, High, Very high].

Fig. 4. Example of GLMP for explaining the classification of beers (left); textual explanation of a classification (right) [8].

models (including fuzzy models) as candidates for generating an explanation. Given an input sample, the simplest interpretable model in accordance with the oracle is used to generate an explanation through LDCP [8].

6 Future Developments

NLG is a promising way for developing XAI systems that generate textual descriptions concerning their inferences. Fuzzy sets seem appropriate models of the meaning of words, therefore fuzzy modeling is a promising approach for NLG, as exemplified by LDCP. Current works are still in the introductory stage and shed light on new research opportunities in the field. In particular, the interaction with deep neural networks is a mid-term objective since it could offer the best of two worlds: the outstanding learning abilities of deep neural networks with the human-centrality of conceptual models like those generated by LDCP.

From the point of view of interpretability in fuzzy modeling, future developments will be focused on representational issues: flat rule-based models are quite standard nowadays but suffer structural limits that could be overcome by more structured representations of knowledge. There are some tentative approaches in this sense by hierarchical fuzzy systems [30] but they are not exempt from criticism [21]. A tighter integration of fuzzy models with explanation models like GLMP may reconcile the need of interpretability of acquired knowledge with the requirement of providing explanation in complex scenarios.

Interpretability itself is matter of ongoing research, in order to cope with current challenges resulting from the higher complexity of data that is used to acquire knowledge. The use of incremental inductive algorithms, for example, is welcome to cope with stream data; nevertheless, these algorithms should take into account the requirement of interpretability of both the resulting knowledge and its historical evolution.

Finally, it must be noticed that the interpretability constraints and criteria, used for an operational definition and assessment of interpretability, are mostly based on common-sense principles. A more formal approach, which looks at interpretability as a protocol for the communication of information semantics, is a promising research direction aimed at establishing the foundations of many methodologies that are under current development.

Acknowledgments. Supported by the Spanish "Ministerio de Economía y Competitividad" through the Ramón y Cajal Program (RYC-2016-19802).

References

1. Alcala-Fdez, J., Alonso, J.M.: A survey of fuzzy systems software: taxonomy, current research trends, and prospects. IEEE Trans. Fuzzy Syst. **24**(1), 40–56 (2016). https://doi.org/10.1109/TFUZZ.2015.2426212
2. Alonso, J.M., Magdalena, L.: Generating understandable and accurate fuzzy rule-based systems in a Java environment. In: Fanelli, A.M., Pedrycz, W., Petrosino, A. (eds.) WILF 2011. LNCS (LNAI), vol. 6857, pp. 212–219. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-23713-3_27

3. Alonso, J., Conde-Clemente, P., Trivino, G.: Linguistic description of complex phenomena with the rLDCP R package. In: Proceedings of the 10th International Conference on Natural Language Generation, pp. 243–244 (2017)
4. Alonso, J.M., Magdalena, L.: HILK++: an interpretability-guided fuzzy modeling methodology for learning readable and comprehensible fuzzy rule-based classifiers. Soft Comput. **15**(10), 1959–1980 (2011). https://doi.org/10.1007/s00500-010-0628-5
5. Alonso, J.M., Magdalena, L., González-Rodríguez, G.: Looking for a good fuzzy system interpretability index: an experimental approach. Int. J. Approx. Reason. **51**(1), 115–134 (2009). https://doi.org/10.1016/j.ijar.2009.09.004
6. Alonso, J.M., Magdalena, L., Guillaume, S.: HILK: a new methodology for designing highly interpretable linguistic knowledge bases using the fuzzy logic formalism. Int. J. Intell. Syst. **23**(7), 761–794 (2008). https://doi.org/10.1002/int.20288
7. Alonso, J.M., Castiello, C., Mencar, C.: Interpretability of fuzzy systems: current research trends and prospects. In: Kacprzyk, J., Pedrycz, W. (eds.) Springer Handbook of Computational Intelligence, pp. 219–237. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-43505-2_14
8. Alonso, J.M., Ramos-soto, A., Castiello, C., Mencar, C.: Hybrid data-expert explainable beer style classifier. In: IJCAI/ECAI Workshop on Explainable Artificial Intelligence (XAI 2018), pp. 1–5 (2018). https://www.dropbox.com/s/jgzkfws41ulkzxl/proceedings.pdf?dl=0
9. Bargiela, A., Pedrycz, W.: Human-Centric Information Processing Through Granular Modelling. SCI, vol. 182. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92916-1
10. Biran, O., Cotton, C.: Explanation and justification in machine learning: a survey. In: Workshop on Explainable AI (XAI), IJCAI 2017, pp. 8–13 (2017). http://www.intelligentrobots.org/files/IJCAI2017/
11. Bustince, H., Barrenechea, E., Fernández, J., Pagola, M., Montero, J.: The origin of fuzzy extensions. In: Kacprzyk, J., Pedrycz, W. (eds.) Springer Handbook of Computational Intelligence, pp. 89–112. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-43505-2_6
12. Casillas, J., Cordón, O., Triguero, F.H., Magdalena, L.: Interpretability Issues in Fuzzy Modeling, vol. 128. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-540-37057-4
13. Castiello, C., Mencar, C., Lucarelli, M., Rothlauf, F.: Efficiency improvement of DC* through a genetic guidance. In: 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–6. IEEE, Naples, July 2017. https://doi.org/10.1109/FUZZ-IEEE.2017.8015585
14. Doran, D., Schulz, S., Besold, T.R.: What does explainable AI really mean? A new conceptualization of perspectives. In: Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017). CEUR Workshop Proceedings, vol. 2071 (2017). http://ceur-ws.org/Vol-2071/CExAIIA_2017_paper_2.pdf
15. Fernandez, A., del Jesus, M.J., Cordon, O., Marcelloni, F., Herrera, F.: Evolutionary fuzzy systems for explainable artificial intelligence: why, when, what for, and where to? IEEE Comput. Intell. Mag., 69–81 (2019). https://doi.org/10.1109/MCI.2018.2881645
16. Gacto, M., Alcalá, R., Herrera, F.: Interpretability of linguistic fuzzy rule-based systems: an overview of interpretability measures. Inf. Sci. **181**(20), 4340–4360 (2011). https://doi.org/10.1016/J.INS.2011.02.021

17. Gatt, A., Krahmer, E.: Survey of the state of the art in natural language generation: core tasks, applications and evaluation. J. Artif. Intell. Res. **61**, 65–170 (2018). https://doi.org/10.1613/jair.5477
18. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–42 (2018). https://doi.org/10.1145/3236009
19. Guillaume, S., Charnomordic, B.: Learning interpretable fuzzy inference systems with FisPro. Inf. Sci. **181**(20), 4409–4427 (2011). https://doi.org/10.1016/J.INS.2011.03.025
20. John, R., Coupland, S.: Type-2 fuzzy logic: challenges and misconceptions [discussion forum]. IEEE Comput. Intell. Mag. **7**(3), 48–52 (2012). https://doi.org/10.1109/MCI.2012.2200632
21. Magdalena, L.: Do hierarchical fuzzy systems really improve interpretability? In: Medina, J., et al. (eds.) IPMU 2018. CCIS, vol. 853, pp. 16–26. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-91473-2_2
22. Mamdani, E.H., Assilian, S.: An experiment in linguistic synthesis with a fuzzy logic controller. Int. J. Man-Mach. Stud. (1975). https://doi.org/10.1016/S0020-7373(75)80002-2
23. Mencar, C., Castiello, C., Cannone, R., Fanelli, A.M.: Design of fuzzy rule-based classifiers with semantic cointension. Inf. Sci. **181**(20), 4361–4377 (2011). https://doi.org/10.1016/j.ins.2011.02.014
24. Mencar, C., Castiello, C., Cannone, R., Fanelli, A.M.: Interpretability assessment of fuzzy knowledge bases: a cointension based approach. Int. J. Approx. Reason. **52**(4), 501–518 (2011). https://doi.org/10.1016/j.ijar.2010.11.007
25. Mencar, C., Fanelli, A.M.: Interpretability constraints for fuzzy information granulation. Inf. Sci. **178**(24), 4585–4618 (2008). https://doi.org/10.1016/j.ins.2008.08.015
26. Mendel, J.: Fuzzy sets for words: a new beginning. In: The 12th IEEE International Conference on Fuzzy Systems, FUZZ 2003, vol. 1, pp. 37–42 (2003). https://doi.org/10.1109/FUZZ.2003.1209334
27. Michalski, R.S.: A theory and methodology of inductive learning. Artif. Intell. **20**, 111–161 (1983). https://doi.org/10.1016/0004-3702(83)90016-4
28. Minsky, M.: Society of Mind. Simon and Schuster, New York (1988)
29. Pinker, S.: How the Mind Works, vol. 882. Wiley/Blackwell (10.1111) (1999). https://doi.org/10.1111/j.1749-6632.1999.tb08538.x
30. Razak, T.R., Garibaldi, J.M., Wagner, C., Pourabdollah, A., Soria, D.: Interpretability indices for hierarchical fuzzy systems. In: Proceedings of IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2017) (2017). https://doi.org/10.1109/FUZZ-IEEE.2017.8015616
31. Revell, T.: Computer says “no comment”. New Sci. **238**(3173), 40–43 (2018). https://doi.org/10.1016/S0262-4079(18)30664-X
32. Schacter, D.L., Gilbert, D.T., Wegner, D.M.: Psychology, 2nd edn. Worth, New York (2011)
33. Sugeno, M., Kang, G.: Structure identification of fuzzy model. Fuzzy Sets Syst. **28**(1), 15–33 (1988). https://doi.org/10.1016/0165-0114(88)90113-3
34. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modeling and control. IEEE Trans. Syst. Man Cybern. **SMC–15**(1), 116–132 (1985). https://doi.org/10.1109/TSMC.1985.6313399
35. Trivino, G., Sugeno, M.: Towards linguistic descriptions of phenomena. Int. J. Approx. Reason. **54**(1), 22–34 (2013). https://doi.org/10.1016/J.IJAR.2012.07.004

36. Wachter, S., Mittelstadt, B., Floridi, L.: Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int. Data Priv. Law **7**(2), 76–99 (2017). https://doi.org/10.1093/idpl/ipx005
37. Wang, Y.: On cognitive informatics. Brain Mind **4**(2), 151–167 (2003). https://doi.org/10.1023/A:1025401527570
38. Yao, Y.: The rise of granular computing. J. Chongqing Univ. Posts Telecommun. Nat. Sci. Ed. **20**(3), 229–308 (2008)
39. Yao, Y.: A triarchic theory of granular computing. Granul. Comput. **1**(2), 145–157 (2016). https://doi.org/10.1007/s41066-015-0011-0
40. Zadeh, L.A.: Information granulation and its centrality in human and machine intelligence. In: 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation, vol. 1, pp. 486–487, October 1997. https://doi.org/10.1109/ICSMC.1997.625798
41. Zadeh, L.A.: From computing with numbers to computing with words. From manipulation of measurements to manipulation of perceptions. IEEE Trans. Circ. Syst. I: Fundam. Theory Appl. **46**(1), 105–119 (1999). https://doi.org/10.1109/81.739259
42. Zadeh, L.A.: A new direction in AI: toward a computational theory of perceptions. AI Mag. **22**(1), 73–84 (2001). https://doi.org/10.1609/aimag.v22i1.1545
43. Zadeh, L.A.: Is there a need for fuzzy logic? Inf. Sci. **178**(13), 2751–2779 (2008). https://doi.org/10.1016/j.ins.2008.02.012
44. Zadeh, L.A.: Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst. **90**(2), 111–127 (1997). https://doi.org/10.1016/S0165-0114(97)00077-8
45. Zhong, N., et al.: Web intelligence meets brain informatics. In: Zhong, N., Liu, J., Yao, Y., Wu, J., Lu, S., Li, K. (eds.) WImBI 2006. LNCS (LNAI), vol. 4845, pp. 1–31. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-77028-2_1