Is Fully Explainable AI Even Possible: Fuzzy-Based Analysis

Miroslav Svitek[a], **Olga Kosheleva**[b] and **Vladik Kreinovich**[c]
[a]Fac. of Transportation Science, Czech Techn. Univ. in Prague, 110 00 Praha 1, Czech Republic,
miroslav.svitek@cvut.cz
[b]Teacher Education, Univ. of Texas at El Paso, El Paso, TX 79968, USA, olgak@utep.edu
[c]Computer Science, Univ. of Texas at El Paso, El Paso, TX 79968, USA, vladik@utep.edu

Abstract

One of the main limitations of many current AI-based decision-making systems is that they do not provide any understandable explanations of how they came up with the produced decision. Taking into account that these systems are not perfect, that their decisions are sometimes far from good, the absence of an explanation makes it difficult to separate good decisions from suspicious ones. Because of this, many researchers are working on making AI explainable. In some applications areas – e.g., in chess – practitioners get an impression that there is a limit to understandability, that some decisions remain *inhuman* – not explainable. In this paper, we use fuzzy techniques to analyze this situation. We show that for relatively simpler systems, explainable model are indeed optimal approximate descriptions, while for more complex systems, there is a limit on the adequacy of explainable models.

Keywords: Explainable AI, Fuzzy logic, Explainability in physics, Interval computations

1 Formulation of the Problem

Good news: deep learning-based systems are very successful. The last decades have seen numerous successes of machine learning – especially deep learning. Systems based on machine learning play chess and Go – and play much better than humans, diagnose some diseases much better than humans, etc.; see, e.g., [4].

Fact: deep learning-based systems are not perfect. Systems based on deep learning are very good, but they are not perfect. There are known examples when a system trained to distinguish, e.g., cats from dogs, mistakenly concludes that the picture is of a dog. Same in more serious applications: deep learning diagnostic systems sometimes misdiagnose a patient, systems for deciding whether to give a loan sometimes make a clearly wrong decision, etc.

Bad news: imperfection of computer-based systems is much more dangerous that imperfection of human decision makers. At first glance, the imperfection of computer-based decision making does not sound so bad: human decision makers also make mistakes, and in many application areas, human decision makers make even more mistakes than computer systems. However, there is a big difference: if you are not sure about the doctor's diagnosis, you can ask the doctor how he/she came up with this diagnosis, and if you (or the second-opinion doctor) do not find it convincing, you can argue against it. In contrast, machine learning-based systems do not provide you with any explanation, so it is difficult to decide which decisions are reasonable and which are based on shaky foundations and need further analysis.

Need for explainable Artificial Intelligence (XAI). In view of the above problem, it is desirable to make AI-based systems explainable.

Natural question. A lot of progress has been done in this direction, but the progress is not as fast as many researchers hoped. So, a natural question can be asked: maybe there is a natural barrier to explainability? maybe fully explainable AI is not possible?

For example, in chess, where computer-based system easily beat up human players, there seems to be a rather general understanding that some moves that the computer systems propose are "inhuman", there is no way to explain them at all – and if someone uses such a move in a game between two humans, this is usually a telltale sign that this person is cheating and using a computer system to help.

What we do in this paper. In this paper, we analyze this question by using fuzzy techniques. Our conclusion is that most probably fully explainable AI is not possible. This does not mean that we should give up on making AI more explainable, it just means that there are natural barriers to explainability.

Comment. This conclusion is similar to a similar situation in complexity theory. It is known that (unless P = NP, which most computer scientists believe to be impossible), no feasible algorithm is possible that would solve all instances of many important problems; see, e.g., [2, 7, 14]. This does not mean that we should not try to solve them, it just means that no matter how much we try, there will always be cases that the current algorithm cannot solve in feasible time.

2 Analysis of the Problem

Let us talk about physics. In order to properly analyze the problem, let us better understand what "explanations" mean for human decision making. And let us take, as an example, an area that is not directly related to potentially emotional issues like illnesses or business success. Let us take an area that does not deal directly with human beings, and does not even deal directly with living beings who may also cause emotions. Let us take an area that deals with the objective world – i.e., physics.

What are explanations in physics: physicists vs. mathematicians. In many areas of physics, we know the equations that describe the corresponding object. In such areas, in solving problems, physicists are helped by mathematicians.

We ourselves have worked with physicists on some such problems, and we were always amazed by the ability of physicists to often solve the corresponding problems faster than we did, in spite of our better mathematical knowledge and our better mathematical skills. How did they do it?

According to the Nobelist Richard Feynman [3], the special skill that physicists have is the ability to find out what can be ignored (at least in the first approximation) and what is important – and thus, to make a complex problem solvable. This leads to a clear and understandable (= explainable) approximate solution. Starting with this solution, we can make it more accurate by taking secondary factors into account, and thus, extend the explanations.

Examples: from Newton to Einstein. This trend can be traced all the way to the first mathematically precise and reasonably universal physical theory – Newton's mechanics. Full equations of celestial mechanics, that take into account the presence of all the planets and their non-zero size, are very difficult to solve even now. They are difficult even if we only take into account the three bodies most importance for us: the Sun, the Earth, and the Moon. What Newton did, in the first approximation, he only considered the Sun and the Earth, he ignored the effects of the Moon and of other planets, and he assumed that both the Sun and the Earth are point bodies – ignoring their size. In this approximation, he could get a clear and understandable solution, and then he showed how to modify it if take Moon into account (this explains tides) and how to take into account non-zero size (this turned out to be automatically taken care of already, at least if we ignore the fact that the Sun and the Earth are not perfect spheres).

This trend can be traced further. There is a (probably apocryphal) story of Einstein's wife invited to the opening of a big research computing center. When she asked what was the purpose of the state-of-the-art computer, she was told – in reference to her husband's work – that one of the main objectives is to study the structure of the Universe. To this she replied that her husband does that on the back of the envelope.

This may be not a true story, but there is another – true – story about Einstein. Many physicists and mathematicians know that it was somewhat accidental that Einstein was the first to come up with equations of General Relativity theory – a relativistic theory of gravitation. Namely, David Hilbert, the most famous mathematician of that time, was also working on this problem, and he came up with the exact same equations – but he submitted his paper 2 weeks later than Einstein. But what many mathematicians do now know is that even if the situation was reversed, and it would have been Einstein who submitted his paper 2 weeks later, physicists would still celebrate mostly Einstein. The reason is very straightforward: all Hilbert did was come with the corresponding very complex and very difficult-to-solve nonlinear system of partial differential equations. Even now high performance computers find solving this system a challenge. Einstein, in his paper, not only came up with these equations. By using a deep understanding of what is important and what can be safely ignored, he came up with simplified equations, solved them, and this way, described possible experimental consequences of this theory. This enabled General Relativity to be experimentally confirmed already in 1919, a few years after it was published.

Summarizing: what seems to be explanations in science. There is a complex phenomenon that is difficult to analyze in all its complexity. To get a good

explainable approximate solution, we select a few features that need to be taken into account, we take them into account fully, and we completely ignore all other features. Correctly selecting the features is not easy, but once this selection is done, we get a simplified system for which we have a clear understandable solution – often even an explicit solution in the analytical form, i.e., described by an explicit formula.

To get a more accurate solution, we can take into account a few more features – again, an important task is to find out which features are most important to add – and we modify the original simple solution to take these additional features into account.

This seems to be an (probably) optimal strategy. Based on the history of physics, this seems to be a successful – thus, probably optimal – strategy. This explains the love for explicit solutions for complex systems of equations, love that may seem to be outdated in modern era where computations are fast – explicit solutions are not only easy to compute, they are also easy to understand, to analyze, and to explain.

But is this indeed an optimal strategy? The fact that for many practical problems, this turned out to be a very successful strategy is an indication that for many such problems, this strategy was indeed probably optimal. On the other hand, the fact that in many new cases, we cannot find such an explanation this way, may mean that for many problems, this strategy is not optimal. So is this strategy always optimal – and if it is not always optimal, why was it optimal for many problems in the past?

To answer these questions, let us formulate this them in precise terms.

3 Towards a Precise Formulation of the Problem and the Resulting Answers

Reminder. Let us denote the number of features we have by n. In the usual physicists' strategy, for each such feature, we either fully take this feature into account, or completely ignore this feature. The overall number of features that we take into account should be small, so that the resulting model will be solvable. Let us denote the overall small number of features that we can take into account by f.

Our main idea: use degrees. In line with Zadeh's main idea that everything is a matter of degree (see, e.g., [1, 6, 10, 12, 13, 17]), we can take into account that for each of n features, in addition to fully taking this feature into account or completely ignoring this feature, we can also take it into account partly, with some degree. Let us denote the degree to which we take this feature into account by d_i. It is reasonable to describe the case when we completely ignore the feature by $d_i = 0$, and the case when we fully take this feature into account by $d_i = 1$. Intermediate cases, when we only partly take the i-th feature into account, can be naturally described by values d_i between 0 and 1.

For degrees, how do we describe the need to limit ourselves to a small amount of information. In situations when each feature is either completely ignored or fully taken into account, for each i, we have either $d_i = 0$ or $d_i = 1$, and the limitations on the total number of features takes the form

$$d_1 + d_2 + \ldots + d_n = f. \tag{1}$$

In situations when we allow partial taking of features into account, to make the problem solvable, we still need to make sure that the overall amount of information that is taken into account is small. A natural way to describe this condition is to similarly require that the sum of all the values d_i is equal to f, i.e., to require the same condition (1).

What do we want. Under the constraint (1), we want to select the values d_i for which the resulting model is as adequate as possible. Let us describe adequacy by a real number: 0 means not adequate, and the larger the value, the higher is the adequacy level. No model is perfectly accurate, so there should not be an upper limit, the values of adequacy should potentially go from 0 to infinity.

The adequacy level a depends on the corresponding degrees: $a = a(d_1, \ldots, d_n)$: if we select the right features, we expect the level of adequacy to be high, but if we select the irrelevant feature, the level of adequacy of the resulting model will be 0.

If we do not select any features at all, the level of adequacy will be 0: $a(0, \ldots, 0) = 0$.

In general, the more information we take into account, the more adequate will be the resulting model. So, the function $a = a(d_1, \ldots, d_n)$ should be non-decreasing in each of its variables: if $d_i < d_i'$, then we should have

$$a(d_1, \ldots, d_{i-1}, d_i, d_{i+1}, \ldots, d_n) \le$$

$$a(d_1, \ldots, d_{i-1}, d_i', d_{i+1}, \ldots, d_n).$$

So, what we want is to find the values d_i that optimize the function $a(d_1, \ldots, d_n)$ under the condition (1). Let us see what we can conclude from this description.

What can we conclude based on such a description: general idea. A priori, we do not know the form of

the function $a(d_1,\ldots,d_n)$. From the purely mathematical viewpoint, it may seem that in this case, we cannot conclude anything definite. However, in physics, such situations – when we do now know the actual function – are common. In such cases, what physicists do (see, e.g., [3, 16]) is take into account that many functions can be expanded in Taylor series

$$f(x_1,\ldots,x_n) = f_0 + \sum_{i=1}^{n} f_i \cdot x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} f_{ij}\cdot x_i \cdot x_j + \ldots$$

for some coefficient f_0, f_i, etc. In particular, $f_0 \stackrel{\text{def}}{=} f(0,\ldots,0)$, each value f_i is the partial derivative of the function $f(x_1\ldots,x_n)$ with respect to x_i at the point $(0,\ldots,0)$, etc.

- In the 0-th approximation, we approximate the function $f(x_1,\ldots,x_n)$ by a constant f_0. If we know the function $f(x_1,\ldots,x_n)$, we can determine the value f_0 as $f_0 = f(0,\ldots,0)$.
- In the first approximation, we take into account linear terms:

  $$f(x_1,\ldots,x_n) = f_0 + \sum_{i=1}^{n} f_i \cdot x_i.$$

  If we know the function $f(x_1,\ldots,x_n)$, then we can determine each value f_i as $f_i = f(0,\ldots,0,1,0,\ldots,0) - f_0$, where 1 is on the i-th place.
- In the next (second) approximation, we also take quadratic terms into account, etc.

Let us apply this general idea to our problem. Let us apply this general idea to our problem. In this case, since $a(0,\ldots,0) = 0$, the 0-th term in the Taylor expansion is equal to 0, so the Taylor expansion has the following form:

$$a(d_1,\ldots,d_n) = \sum_{i=1}^{n} a_i \cdot d_i + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\cdot d_i \cdot d_j + \ldots$$

Simplest case explains the classical explainable strategy. In the simplest case, we only take into account linear terms. In this case, we need to optimize the function

$$a(d_1,\ldots,d_n) = \sum_{i=1}^{n} a_i \cdot d_i \qquad (2)$$

under the constraint (1). Here, the objective function is linear, and the set fo all possible values of the variables d_i – as determined by linear inequalities $0 \le d_i$, $d_i \le 1$, and (1) – is a convex polyhedron. It is known (see, e.g., [15]) that on each convex polyhedron, a linear function attains its maximum on one of the vertices, i.e., in one of the tuples in which n inequalities become equalities.

So, in addition to equality (1), $n-1$ inequalities must become equalities. For each i, we cannot have two equalities: this would mean that d_i is equal both to 0 and to 1. So, for each i, we can only have one valid equality. Thus, for $n-1$ values i, we must have $d_i = 0$ or $d_i = 1$. The remaining value d_{i_0} can be determined by the condition (1), as

$$d_{i_0} = f - (d_1 + \ldots + d_{i_0-1} + d_{i_0+1} + \ldots + d_n). \qquad (3)$$

All the values f and d_i in the right-hand side of the formula (3) are integers, so the resulting value d_{i_0} is also an integer. Since each degree d_i is located in the interval $[0,1]$, the only two possible choices for the integer degree d_{i_0} are $d_{i_0} = 0$ or $d_{i_0} = 1$. So, for each i, we will have $d_0 = 0$ or $d_i = 1$, i.e., we will have a classical explainable solution – which is thus shown to be the actually optimal one.

Comment. It this case, it is also easy to describe for which exactly values d_i the maximum of the expression (2) is attained., For this purpose, we sort the values a_i in decreasing order:

$$a_{(1)} \ge a_{(2)} \ge \ldots \ge a_{(n)},$$

and then select $d_i = 1$ for the values i that correspond to f largest values of a_i:

$$d_{(1)} = d_{(2)} = \ldots = d_{(f)} = 1$$

and $d_i = 0$ for all other i.

What if the situation becomes more complex. When the situation becomes more complex, it is no longer possible to only use a linear approximation, we will have to use the quadratic approximation instead:

$$a(d_1,\ldots,d_n) = \sum_{i=1}^{n} a_i \cdot d_i + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\cdot d_i \cdot d_j. \qquad (4)$$

In this case, we face the problem of maximizing the quadratic function (4) under the linear constraint (1) and the constraints $0 \le d_i \le 1$.

For the linear function, the maximum is always attained at the tuple $(d_1,\ldots,d_n)$ for which each value d_i is 0 or 1 – and, if we know the function $a(d_1,\ldots,d_n)$, it is easy to find this optimal tuple. In contrast, in the quadratic case, the maximum may be attained for values $d_i \in (0,1)$ and, what is even worse, even when we know the quadratic function, finding the values d_i at which this function attains maximum is, in general, NP-hard; see, e.g., [7].

In such more complex situations, the traditional explanation approach is no longer optimal – and finding the appropriate explanation is an NP-hard problem – which means that the problem is no longer feasible.

Comment. It should be mentioned that our problem of maximizing a function under constraints $d_i \in [0,1]$ is a particular case of the general problem of maximizing or minimizing a function $f(x_1,\ldots,x_n)$ under the interval constraints $x_i \in [\underline{x}_i,\overline{x}_i]$, the problem studied in *interval computations*; see, e.g., [5, 8, 9, 11].

Conclusions. When the systems that we study are sufficiently simple, so that a linear approximation reasonably accurately describes the adequacy level, the optimal way to provide a simplified model is to fully take into account a small number of features, and to ignore all other features. This is exactly what the traditional explainable models do.

However, as the systems become more complex, the traditional explainable approach is no longer optimal. In an optimal approximation, we need to take into account many features — to some degree. This is what we observe in finite-element solutions to partial differential equations, this is what we observe in machine learning – we have adequate models, but these models are not explainable in the original sense of this word.

Thus, our conclusion is that, as the systems that we analyze become more and more complex, there is a limit of explainability, we will have to live with the fact that some conclusions of these models are "inhuman".

4 Auxiliary Idea: What if We Take Uncertainty Into Account

Idea and how to describe it. In the previous sections, we considered well-defined models, in which some features are taken into account (maybe to a degree) and some features are ignored. Such models provide an approximate description of the system, but by themselves, they do not provide us with an idea of how accurate is this description.

A natural way to provide this additional description is, e.g., instead of simply ignoring a feature, to take into account that this feature is present – without specifying what exactly is the effect of this feature. Since we decided to describe ignoring a feature by 0 and fully taking it into account by 1, it is natural to describe this new idea – where we do not know what will be the actual degree of this feature's effect – by the whole interval $[0,1]$.

In general, if we explicitly take into account the i-th feature with some degree $\underline{d}_i$, and we also take into account the possibility that this feature may have an effect corresponding to a higher degree $\overline{d}_i$, then it is reasonable to describe this situation by an interval $[\underline{d}_i,\overline{d}_i]$.

Comment. While the original degrees $d_i \in [0,1]$ corresponded to the usual fuzzy logic, the above-mentioned degrees $[\underline{d}_i,\overline{d}_i]$ correspond to interval-valued fuzzy logic; see, e.g., [10].

Let us formulate this idea in precise terms. For each i, instead of a single value d_i, we select an interval $[\underline{d}_i,\overline{d}_i]$. In each selection, we absolutely take into account each feature to the degree $\underline{d}_i$. The overall amount of features that we absolutely take into account should be equal to f:

$$\underline{d}_1 + \ldots + \underline{d}_n = f. \tag{5}$$

The amount of uncertainty corresponding to each feature is equal to the difference $\overline{d}_i - \underline{d}_i$: when this difference is 0, we are back to the previous scheme that does not take uncertainty into account at all. Taking uncertainty into account is not easy. So, the overall amount of uncertainty that we can take into account should also be limited by some number u:

$$(\overline{d}_1 - \underline{d}_1) + \ldots + (\overline{d}_n - \underline{d}_n) = u. \tag{6}$$

What is the optimal strategy: simplest case. Under the constraints (5) and (6), we need to maximize the adequacy, which now depends on both bounds $\underline{d}_i$ and $\overline{d}_i$. In the simplest case, the adequacy is described by a linear function

$$a(\underline{d}_1,\overline{d}_1,\ldots,\underline{d}_n,\overline{d}_n) = \sum_{i=1}^{n} (\underline{a}_i \cdot \underline{d}_i + \overline{a}_i \cdot \overline{d}_i).$$

So, we need to maximize this function under the conditions (5), (6), and

$$0 \le \underline{d}_i \le \overline{d}_i \le 1 \tag{7}$$

Similarly to the previous case, the maximum is attained when $2n$ inequalities become equalities. In each system (7), at most two inequalities can becomes equalities. The case when exactly two become equalities is when the interval $[\underline{d}_i,\overline{d}_i]$ is equal to $[0,0]$, to $[1,1]$, or to $[0,1]$. If this happens for fewer than $n-2$ values i, then we will have:

- fewer than $2(n-2) = 2n-4$ such equalities
- plus two equalities (5) and (6)

- plus two equalities from the two remaining indices i,

to the total of less than $2n$. Since we do have $2n$ equalities, this implies that for at least $n-2$ indices (i.e., for almost all indices), we should have either 0 (ignoring this feature), or 1 (fully taking this feature into account), or $[0,1]$ – meaning that we fully take into account the *possibility* of this feature's effect, without providing any details about the actual effect.

This is exactly what physicists do. The above conclusion is also in line with what physicists do: to estimate uncertainty, they usually fully take into account some sources of uncertainty while ignoring other sources.

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), HRD-1834620 and HRD-2034030 (CAHSI Includes), EAR-2225395, and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

The authors are greatly thankful to Yaacov Kopeliovich for useful discussion about the importance of exact solutions, and to the anonymous referees for valuable suggestions.

References

[1] R. Belohlavek, J. W. Dauben, and G. J. Klir, *Fuzzy Logic and Mathematics: A Historical Perspective*, Oxford University Press, New York, 2017.

[2] Th. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, Massachusetts, 2022.

[3] R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison Wesley, Boston, Massachusetts, 2005.

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, Massachusetts, 2016.

[5] L. Jaulin, M. Kiefer, O. Didrit, and E. Walter, *Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control, and Robotics*, Springer, London, 2001.

[6] G. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic*, Prentice Hall, Upper Saddle River, New Jersey, 1995.

[7] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1998.

[8] B. J. Kubica, *Interval Methods for Solving Nonlinear Constraint Satisfaction, Optimization, and Similar Problems: from Inequalities Systems to Game Solutions*, Springer, Cham, Switzerland, 2019.

[9] G. Mayer, *Interval Analysis and Automatic Result Verification*, de Gruyter, Berlin, 2017.

[10] J. M. Mendel, *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*, Springer, Cham, Switzerland, 2017.

[11] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM, Philadelphia, 2009.

[12] H. T. Nguyen, C. L. Walker, and E. A. Walker, *A First Course in Fuzzy Logic*, Chapman and Hall/CRC, Boca Raton, Florida, 2019.

[13] V. Novák, I. Perfilieva, and J. Močkoř, *Mathematical Principles of Fuzzy Logic*, Kluwer, Boston, Dordrecht, 1999.

[14] C. Papadimitriou, *Computational Complexity*, Addison-Wesley, Reading, Massachusetts, 1994.

[15] R. T. Rockafeller, *Convex Analysis*, Princeton University Press, Princeton, New Jersey, 1997.

[16] K. S. Thorne and R. D. Blandford, *Modern Classical Physics: Optics, Fluids, Plasmas, Elasticity, Relativity, and Statistical Physics*, Princeton University Press, Princeton, New Jersey, 2021.

[17] L. A. Zadeh, "Fuzzy sets", *Information and Control*, 1965, Vol. 8, pp. 338–353.