Evaluating Explanations by Cognitive Value

Ajay Chander, Ramya Srinivasan

Evaluating Explanations by Cognitive Value

Ajay Chander and Ramya Srinivasan

Fujitsu Laboratories of America, Sunnyvale, CA, 94085, USA

Abstract. The transparent AI initiative has ignited several academic and industrial endeavors and produced some impressive technologies and results thus far. Many state-of-the-art methods provide explanations that mostly target the needs of AI engineers. However, there is very little work on providing explanations that support the needs of business owners, software developers, and consumers who all play significant roles in the service development and use cycle. By considering the overall context in which an explanation is presented, including the role played by the human-in-the-loop, we can hope to craft effective explanations. In this paper, we introduce the notion of the "cognitive value" of an explanation and describe its role in providing effective explanations within a given context. Specifically, we consider the scenario of a business owner seeking to improve sales of their product, and compare explanations provided by some existing interpretable machine learning algorithms (random forests, scalable Bayesian Rules, causal models) in terms of the cognitive value they offer to the business owner. We hope that our work will foster future research in the field of transparent AI to incorporate the cognitive value of explanations in crafting and evaluating explanations.

Keywords: Explanations · AI · Cognitive Value· Business Owner · Causal Modeling.

1 Introduction

Consumers, policymakers, and technologists are becoming increasingly concerned about AI as a 'black-box' technology. In order to engender trust in the user and facilitate comfortable interactions, it has become increasingly important to create AI systems that can explain their decisions to their users. Across a variety of fields, from healthcare to education to law enforcement and policy making, there exists a need for explaining the decisions of AI systems. In response to this, both the scientific and industrial communities have shown a growing interest in making AI technologies more transparent. The new European General Data Protection Regulation, the U.S. Defense Advanced Research Projects Agency's XAI program [1], and institutional initiatives to ensure the safe development of AI such as those of the Future of Life Institute, are a few of the many business, research, and regulatory incentives being created to make AI systems more transparent.

Many state-of-the-art methods provide explanations that mostly target the needs of AI engineers [10, 11, 13]. In other words, explanations assume some domain knowledge, or are generated for people with domain expertise. As the use of

AI becomes widespread, there is an increasing need for creating AI systems that can explain their decisions to a large community of users who are not necessarily domain experts. These users could include software engineers, business owners, and end-users. By considering the overall context in which an explanation is presented, including the role played by the human-in-the-loop, we can hope to craft effective explanations.

1.1 Cognitive Value of an Explanation

The role of explanations and the way they should be structured is not new and dates back to the time of Aristotle [25]. The authors in [25] highlight the functions of explanations. They mention that explanations should accommodate novel information in the context of prior beliefs, and do so in a way that fosters generalization. Furthermore, researchers have also studied if certain structures of an explanation are inherently more appealing than others [26]. The authors in [23] state that explanations are social in that they are meant to transfer knowledge, presented as part of a conversation or interaction and are thus presented relative to the explainer's beliefs about the user's (i.e., explainee's) beliefs.

We posit that *an explanation is a filter on facts*, and is presented and consumed as part of a larger context. Here, fundamental aspects of the context include: the entity presenting the explanation ("explainer"), the entity consuming the explanation ("explainee"), the content of the explanation itself, where the explanation is being presented, amongst others.

Let's first understand the role of the explainee as it is the most crucial element of an explanation's context. As discussed earlier, a wide variety of users are now interested in understanding the decisions of AI systems. There are at least four distinct kinds of users [3, 4].

- *AI Engineers:* These are generally people who have knowledge about the mathematical theories and principles of various AI models. These people are interested in explanations of a functional nature, e.g. the effects of various hyperparameters on the performance of the network or methods that can be used for model debugging.
- *Software Developers and/or Integrators:* These are application builders who make software solutions. These users often make use of off-the-shelf AI modules, and integrate them with various software components. Developers are interested in explanation methods that allow them to seamlessly integrate various AI module into the use cases of their interest.
- *Business Owners:* These people are usually stakeholders who own the service and are interested in commercialization. The owner is concerned with explainability aspects that can elucidate ways in which the application can be improved to increase financial gains, to justify predictions in order to aid in product design and people management, etc.
- *End-Users:* These are consumers of the AI service. These people are interested in understanding why certain recommendations were made, how they can use the information provided by the AI, how the recommendations will benefit them, etc.

As described above, users expect certain "cognitive values" from the explanations of AI systems. The term cognitive value can be best explained via examples. Some users may primarily expect explanations to account for personal values (e.g., privacy, safety, etc.) in the recommendations made by AI systems. In this case, the cognitive value of the explanation is to *engender trust* in the user. Some other users may largely expect explanations to be elucidating functional aspects of the AI models such as accuracy, speed and robustness; here the cognitive value of explanation is in aiding *troubleshooting and/or design*. Some users may expect explanations to help them understand the AI's recommendation and aid them in analysis; in this case the cognitive value of explanation is in *educating* the user and help them take an appropriate *action*. Based on the task, any of the aforementioned cognitive values may be important to any of the user-types described. There could be many more cognitive values, but we believe that *trust, troubleshooting, design, education and action* are the most important cognitive values.

Thus, it becomes important to evaluate explanations based on their cognitive value in a given context. As an example, consider a business executive who wants to understand how to improve sales of the company. So, the operational goals of the explanation is largely in aiding *action* (i.e., the AI should help the business executive in specifying the steps that need to be taken in order to improve sales) and in *education* (i.e., the AI should inform the executive of the factors that determine sales, etc.). Consider some hypothetical explanations generated by an AI system as listed below.

- Factors X and Y are the most important factors in determining sales
- Factors X and Y are the most important factors in determining sales, whenever $X > 5$ and $Y < 4$, the sales is 90%.
- Factors X and Y are the most important factors responsible for sales in the past. Changing X to X+10 will improve the sales by 5%.

At a high-level, all of the aforementioned explanations look reasonable. Let us delve a little deeper. Suppose X was the amount of the product and Y was the location of the sale. Now, in explanation 2, the phrase "$Y < 4$" does not convey a semantic meaning to the business owner. To the AI engineer, it may be still meaningful as the model might have mapped various locations to numbers. However, the business owner is not aware about this encoding. Even if she was made aware of what the individual numbers denoted (such as if the location is NYC, Tokyo, or Hamburg), as the number of such choices increases, the cognitive burden on the business owner increases and does not aid in educating him/her or aiding in their action of how they can improve sales. Although explanation 1 provides semantically relevant information, it does not help the business owner in providing actionable insights in improving the sales. Explanation 3 not only educates the business owner in terms of the most important factors for improving sales, but more importantly also aids in action by suggesting *how* the sales can be improved.

The *contributions* of the paper are as follows: First, to the best of our knowledge, our work is the first to introduce the notion of "cognitive value" of an ex-

planation and elaborate on the role of cognitive values in providing explanations to various kinds of users. Second, we compare three state-of-the-art explanation methods namely Scalable Bayesian Rule Lists [7], Random Forests, and Causal models [5] in terms of their cognitive value to the business owner. In particular, through a case study of a car dealer who is wanting to improve car sales, we show how causal models designed for explaining issues concerning fairness and discrimination [5] can be modified to provide explanations of cognitive value to this car dealer. Third, we discuss the merits and shortcomings of each of the aforementioned methods. We hope that our work will foster future research in the field of transparent AI to incorporate the cognitive value of explanations in evaluating the AI-generated explanations.

The rest of the paper is organized as follows. An overview of related work is provided in Section 2. The case study and the dataset is described in Section 3. Section 4 provides background on causal models, scalable bayesian rule lists and random forest algorithms. It also includes a description of how the causal model proposed in [5] for detecting discrimination can be leveraged to provide explanations of cognitive value. The types of explanations obtained from the three models are summarized in Section 5. A discussion of the relative merits and shortcomings of the explanations obtained by each of the aforementioned methods is also provided in Section 5. Conclusions are provided in Section 6.

2 Related Work

The new European General Data Protection Regulation (GDPR and ISO/IEC 27001) and the U.S. Defense Advanced Research Projects Agency's XAI program [1] are probably the most important initiatives towards transparent AI. As a result, several academic as well as industrial groups are looking to address issues concerning AI transparency. Subsequently, a series of workshops, industrial meetings and discussion panels related to the area have taken place contributing to some impressive results.

Most of the work in the area is oriented towards the AI engineer and is technical. For example, in [10], the authors highlight the regions in an image that were most important to the AI in classifying it. However, such explanations are not useful to an end-user in either understanding the AI's decision or in debugging the model [14]. In [19], the authors discuss the main factors used by the AI system in arriving at a certain decision and also discuss how changing a factor changes the decision. This kind of explanation helps in debugging for the AI engineers. Researchers are also expanding the scope of explanations to AI agents by proposing frameworks wherein an AI agent explains its behavior to its supervisor [27]. The authors in [28] propose a model agnostic explanation framework and has been instrumental in several subsequent research efforts. There are several other impressive works across various fields catered towards helping the AI engineers [11, 13, 15, 16, 29–32, 40, 38]. A nice summary concerning explainability from an AI engineer's perspective is provided in [22, 34].

More recently, there have been efforts in understanding the human interpretability of AI systems. The authors in [24] provide a taxonomy for human interpretability of AI systems. A nice non-AI engineer perspective regarding explanations of AI system is provided in [23]. The authors in [17] studied how explanations are related to user trust. They conducted a user study on healthcare professionals in AI-assisted clinical decision systems to validate their hypotheses. A nice perspective of user-centered explanations is provided in [18]. The author emphasizes the need for persuasive explanations. The authors in [21, 36] explore the notion of interactivity from the lens of the user. With growing interest in the concept of interpretability, various measures for quantifying interpretability have also been proposed in [40, 41, 39].

The closest to our work is perhaps [20] wherein the authors discuss how humans understand explanations from machine learning systems through a user-study. The metrics used to measure human interpretability are those concerning explanation length, number of concepts used in the explanation, and the number of repeated terms. Interpretability is measured in terms of the time to response and the accuracy of the response. Our measure is on the cognitive value an explanation offers as opposed to time to response or other such quantitative measures.

3 Case Study and Dataset

Our focus is on non-AI engineers. As a case study, we consider a scenario involving a business owner. Specifically, we consider a car dealer who wants to improve the sales of the cars. Thus, this user will benefit from knowing the steps that need to be taken in order to increase the sales of the cars. Thus, the cognitive value an explanation offers in this scenario should be in guiding towards an appropriate action and justifying the same.

We consider the car evaluation dataset [28, 42, 43] for our analysis, obtained from the UCI Machine learning repository. Although relatively an old dataset, it is appropriate for the problem at hand. The dataset is a collection of six attributes of cars as listed in Table 1. In the original dataset, the output attributes are "acceptable", "unacceptable", "good", and "very good". For the specific case study considered, we map acceptance to sales. For evaluation purposes, we map probability of acceptability to probability of selling the car and probability of unacceptability to probability of not being able to sell the car. There are 1728 instances in the dataset. The car dealer is interested in knowing what changes need to be done i.e., what factors of the cars need to be changed in order to improve sales.

4 Background

We consider three state-of-the-art algorithms for comparing the cognitive value they offer in the context of the aforementioned case study. We consider Random forests [8] as this model is one of the earliest interpretable models. We also

Table 1. Description of input variables in the Car Evaluation Dataset.

Input Variable	Values
Buying price	vhigh, high, med, low
Price of the maintenance	vhigh, high, med, low
Number of doors	2, 3, 4, 5more
Persons capacity in terms of persons to carry	2, 4, more
Size of luggage boot	small, med, big
Estimated safety of the car	low, med, high

consider two recent models scalable Bayesian rules [7] proposed in 2017, and causal models [5] proposed in 2018. For completeness, we provide some basic background about these models in the context of interpretability.

4.1 Random Forests

Random forests are a class of ensemble learning methods for classification and regression tasks [8]. The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, ..., x_n$ with labels $Y = y_1, ..., y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples, i.e.,
For $b = 1, ..., B$:
1. Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.
2. Train a classification or regression tree f_b on X_b, Y_b.
After training, predictions for unseen samples x's can be made by averaging the predictions from all the individual regression trees on x's or by taking a majority vote in the case of classification trees.

When considering a decision tree, for each decision that a tree (or a forest) makes there is a path (or paths) from the root of the tree to the leaf, consisting of a series of decisions, guarded by a particular feature, each of which contribute to the final predictions. The decision function returns the probability value at the leaf nodes of the tree and the importance of individual input variables can be captured in terms of various metrics such as the Gini impurity.

4.2 Scalable Bayesian Rule Lists (SBRL)

SBRLs are a competitor for decision tree and rule learning algorithms in terms of accuracy, interpretability, and computational speed [7]. Decision tree algorithms are constructed using greedy splitting from the top down. They also use greedy pruning of nodes. They do not globally optimize any function, instead they are composed entirely of local optimization heuristics. If the algorithm makes a mistake in the splitting near the top of the tree, it is difficult to undo it, and consequently the trees become long and uninterpretable, unless they are heavily pruned, in which case accuracy suffers [7]. SBRLs overcome these shortcomings of decision trees.

Bayesian Rule Lists is an associative classification method, in the sense that the antecedents are first mined from the database, and then the set of rules and their order are learned. The rule mining step is fast, and there are fast parallel implementations available. The training set is $(x_i, y_i)_i^n$, where the $x_i \in X$ encode features, and y_i are labels, which are generally either 0 or 1. The antecedents are conditions on the x that are either true or false. For instance, an antecedent could be: if x is a patient, antecedent a_j is true when the value of x is greater than 60 years and x has diabetes, otherwise false. Scalable Bayesian Rule Lists maximizes the posterior distribution of the Bayesian Rule Lists algorithm by using a Markov Chain Monte Carlo method. We refer interested readers to [7] for greater details related to the working of the algorithm.

4.3 Causal Models

Causal models are amenable towards providing explanations as they naturally uncover the cause-effect relationship [37]. Before describing how causal models can be used to elicit explanations, we list some basic definitions used.

Terminologies: A structural causal model (SCM) M is a tuple $h = [U, V, F, P(U)_i]$ where: U is a set of exogenous (unobserved) variables, which are determined by factors outside of the model; V is a set $V_1, ..., V_n$ of endogenous (observed) variables that are determined by variables in the model; F is a set of structural functions $f_1, ..., f_n$ where each f_i is a process by which V_i is assigned a value v_i; $P(u)$ is a distribution over the exogenous variables U [5].

Each SCM M is associated with a causal diagram G, which is a directed acyclic graph where nodes correspond to the endogenous variables (V) and the directed edges denote the functional relationships. An intervention, denoted by $do(X = x)$ [9], represents a model manipulation where the values of a set of variables X are set fixed to x regardless of how their values are ordinarily determined (f_x). The counterfactual distribution is represented by $P(Y_{X=x} = y)$ denotes the causal effect of the intervention $do(X = x)$ on the outcome Y , where the counterfactual variable $Y_{X=x}$ (Y_x , for short) denotes the potential response of Y to intervention $do(X = x)$. We will consistently use the abbreviation $P(y_x)$ for the probabilities $P(Y_{X=x} = y)$, so does $P(y|x) = P(Y = y|X = x)$.

For our analysis, we consider a standard model provided in [5] as depicted in Figure 1. We wish to determine the effect of X on Y (say X= safety and Y= car sales). In this context, X would be the input factor and Y would be the output factor. There could be other factors Z and W affecting sales as shown in Figure 1. Here the factor Z is a common cause and is often referred to as a confounder. The factor W is called a mediator, because X could have a causal effect on Y through W.

There are three types of causal effects defined with respect to Fig. 1. The direct effect is modeled by the direct causal path $X \to Y$ in Fig. 1. Indirect effect is modeled by the path $X \to W \to Y$ and spurious effect is jointly modeled by the paths $Z \to X$ and $Z \to Y$.

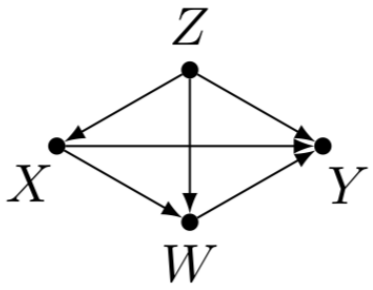

Fig. 1. Structural Causal Model considered for our analysis.

For each SCM, one can obtain the direct, indirect and spurious effects of X on Y. In particular, the authors in [5] define the concepts of counterfactual direct effect, counterfactual indirect effect and counterfactual spurious effects in order to estimate the discover discrimination and argue that by disentangling each of the causal effects, it can be ascertained whether there was genuine discrimination or not. The direct, (D.E.) indirect (I.E) and spurious (S.E.) causal effects of changing the various factors on the output can be obtained from the following equations as provided in [5]. For more elaborate details, we refer readers to [5].

$$D.E_{x_0,x_1}(y|x) = \sum_{z,w}((P(y|x_1,w,z) - P(y|x_0,w,z))P(w|x_0,z)P(z|x) \tag{1}$$

$$I.E_{x_0,x_1}(y|x) = \sum_{z,w}(P(y|x_0,w,z)(P(w|x_1,z) - P(w|x_0,z))P(z|x) \tag{2}$$

$$S.E_{x_0,x_1}(y) = \sum_{z,w}(P(y|x_0,w,z)P(w|x_0,z(P(z|x_1) - P(z|x_0)) \tag{3}$$

4.4 Adaptation of Causal Models to provide Cognitive Value:

Although the purpose of the authors of [5] was to explain discrimination, it is straightforward to extend this to obtain explanations that can offer cognitive values. Below, we describe the steps that need to be followed to obtain explanations of cognitive value.

Put in other words, we consider all possible SCMs for the choice of factors $[X, Z, W]$ as input, mediator and confounder. Note, for the standard model considered, only one confounder and one mediator is allowed. For the car evaluation dataset, we consider 4 factors for each SCM. Let us understand the above process for the car evaluation dataset.

1 Estimate the counterfactual direct effects for all possible combinations of SCMs for a given input X and output Y.
2 Repeat Step 1 for all possible choice of input factors X.
3 For each choice of input factor, generate textual statements highlighting the differential probability in output (e.g. differential probability in selling car) for change in the value of the input factor (e.g. changing the safety of the car from low to high).
4 The factors corresponding to highest differential probabilities offer the most cognitive value (i.e. to increase sales) to the user (e.g. a car dealer).

Let us understand the usability of the aforementioned algorithm for the case study considered. We are interested in explaining how to improve the sales to the business owner (who is the car dealer in this example). So, the factor Y corresponds to sales. Suppose, X =safety and Y= sale. In the model shown in Fig 1, one possibility could be W = number of persons and Z could be maintenance. This means, safety could affect car sales through the factor number of persons, and the factor maintenance could be a confounder affecting both safety and sales. Another possibility could be that W = maintenance and Z could be number of persons. In this case, the factor number of persons is a confounder and affects both sales and safety, and maintenance is a mediator.

Let us first consider the case wherein X is safety, Z is maintenance and let W be number of persons. Putting this in the standard model of Fig.1 and using Eq. 1, we can estimate the counterfactual direct effect of safety on sales. The concept of counterfactual direct effect can be best understood through an example. Suppose there is a car with low safety. All other factors unchanged, if the factor safety alone were to be changed to high, then the quantity "counterfactual direct effect" can provide a measure of the improvement in sales for this factor change. Please note, in reality, since all the cars are manufactured, none of the factors can be changed. But, for the time being, assume an imaginary car whose factors can be changed. In that scenario, if the safety of the imaginary car were to be high, then one can ascertain if that change in safety contributes to rise or fall of sales and by how much. Knowing this differential sales probability will help in future design of such cars for the car dealer. Thus, it provides the cognitive value in taking appropriate *action* to the car dealer. We compute counterfactual direct effect for all possible choices of input factors X. Since the output factor is the sales, we conclude that factors with the highest magnitude of the counterfactual direct effect are the most important ones for the car dealer in improving the sales.

5 Dataset Analysis and Results

In this section, we state the results obtained from each of the three methods discussed in Section 4. We compare the three methods in terms of their cognitive value to the car dealer.

5.1 Results from Random Forests

The algorithm returns the probability value of sales for individual cars. In addition, variable importance scores in terms of mean decreasing impurity is provided that explains the importance of individual factors (i.e. safety,number of persons, etc.) in determining the sale of a car. Table 2 lists the variable importance scores in terms of mean decreasing Gini.

Table 2. Results from Random Forest Algorithm.

Input Factor	Importance
Buying price	92.15
Price of the maintenance	97.36
Number of doors	27.86
Persons capacity in terms of persons to carry	178.52
Size of luggage boot	51.19
Estimated safety of the car	215.87

It is apparent from the above table that safety and number of persons that can be accommodated are the most important factors in determining the sales of the cars. This information can certainly educate the car dealer about the most important factors that determine the sales.

5.2 SBRL

Let us next consider the result from scalable Bayes Rules List. As stated earlier, it is in the form of "if-then" associative rules. The results for the car evaluation dataset is as follows. Please note, the phrase 'positive probability' refers to the sale of the car. The rule numbers are generated by the algorithm and simply refer to the condition mentioned beside it in text. For example, rule [105] refers to the condition 'number of persons =2'.

If [persons=2] (rule[105]) then positive probability = 0.00173010
else if [safety=low] (rule[124]) then positive probability = 0.00259067
else if [doors=2,lug-boot=small] (rule[31]) then positive probability = 0.28787879
else if [buying=med,safety=high] (rule[22]) then positive probability = 0.98888889
else if [buying=low] (rule[16]) then positive probability = 0.94382022
else if [maint=vhigh,lug-boot=small] (rule[94]) then positive probability = 0.03125000
else if [buying=med] (rule[25]) then positive probability = 0.84523810
else if [lug-boot=small,safety=med] (rule[68]) then positive probability = 0.02631579
else if [maint=vhigh] (rule[98]) then positive probability = 0.01515152
else if [lug-boot=med,safety=med] (rule[64]) then positive probability = 0.52000000
else if [buying=high] (rule[10]) then positive probability = 0.98913043
else if [maint=high] (rule[77]) then positive probability = 0.03125000

Thus, SBRLs provide various conditions and state the probability of sales under that condition. Thus, if the number of persons is 2, the probability of sales is 0.1%.

5.3 Causal Models

Table 3 provides a summary of the results obtained from the causal model described in Section 4.4.

Table 3. Results from Causal Explanation.

Input Factor	Real Car	Imaginary Car	Differential Probabilities (expressed as %) in selling Real car- selling Imaginary car
Safety	Low	High	+36.36%
Safety	High	Low	-50%
Number of persons	2	4	+32.34%
Number of persons	4	2	-43%
Maintenance	High	Low	+2.5%
Maintenance	Low	High	-10%

The results summarized in Table 3 can be understood via examples. As an instance, consider the first row corresponding to safety. The result of that row states - *"All other factors unchanged, if the safety of the car is changed from low to high, there will be 36.36% improvement in sales.* The next row corresponding to safety reads thus : "*All other factors unchanged, if the safety of the car is changed from high to low, there will be 50% drop in sales."* . A positive value of differential probability indicates that there will improvement in sales upon changing the corresponding input factor (e.g. sales) in the stated manner (i.e. from low to high safety). A negative differential probability corresponds to a drop in sales.

Table 3 re-iterates the result of random forest. It can be noted that safety and number of persons are the most important factors in determining sales. Note, Table 3 highlights the most important factors in improving sales and hence some factors (e.g. lug-boot) are omitted from the table.

5.4 Discussion

In this section, we discuss the merits and de-merits of all the three methods from the perspective of cognitive value the respective explanations offer to the users

Random Forest: The random forest educates the car dealer about the most important factors responsible for car sales in a relative manner. The significance of absolute values of the importance scores is not clear as their range is unknown. Furthermore, knowing the factor importance scores does not help the car dealer in understanding what needs to be done in order to improve the sales. The result may thus only educate the car dealer in knowing the most important factors

affecting sales, but it is unclear as to how those factors need to be changed in order to improve sales.

SBRLs: There are many if-else statements in the explanation provided by SBRLs. The specific conditions are chosen automatically by the algorithm and can consist of multiple conditions that may be difficult for the user to interpret. Even if one parses for the ones with highest positive probabilities (implying sales of cars), it neither conveys semantically relevant information nor provides actionable insights to the car dealer. For example, the highest probability of 0.989 corresponds to the rule "if buying= high". Does this mean cars with high buying price sell more? If true, it does not seem practically very true or compelling. Even assuming that it is true, it does not provide actionable insights to the car owner. By how much can the price be increased to achieve a certain sales target? Such kind of information is lacking in this model's result.

Causal Models: Unlike random forest which could not ascertain how those factors are important and in particular how the car dealer should change those to improve sales, the explanation from the causal model provides *actionable* insights to the car dealer in improving sales. Furthermore, the results from the causal model is semantically meaningful and practically relevant.

Although the explanations based on causal modeling offers cognitive value to the users, it comes at a price. Specifically, one has to try with different structural assumptions. For a non-domain expert, this can really be time consuming. Also, the causal explanation formula works best for binary data. While this is good in providing instance level explanations (local explanations), it may not be easy to derive global explanations.

Table 4 provides a comparison of the three methods in terms of their cognitive value to the car dealer.

Table 4. Comparison of Results: RF denotes random forests, CM denotes causal models, SBRL denotes Scalable Bayesian Rule Lists

Method	Educates the user	Provides Actionable insights to the user	Easy to comprehend
RF	provides relative importance of factors in sales	No	Range of variable importance is not clear
SBRL	Informs about sales under certain conditions	No	several conditions to parse
CM	provides relative importance of factors in sales	explains how the sales can be improved	Yes

6 Conclusions

We introduced the concept of "cognitive value" of explanations of AI systems to users. We considered the scenario of a business owner seeking to improve sales

of their product and compared explanations provided by some state-of-the-art AI methods in terms of the cognitive value they offer to the business owner. Specifically, we studied random forest, scalable bayesian rule lists and causal explanations towards this end. For the context considered, causal explanations provided the best cognitive value in terms of providing the business owner with actionable insights in improving his/her sales. We hope our work will foster future research in the field of transparent AI to incorporate the cognitive value of explanations in assessing explanations.

References

1. Gunning, D. : Explainable artificial intelligence (XAI). Defense Advanced Research Projects Agency (DARPA), 2017
2. Miller, T. : Explanation in Artificial Intelligence: Insights from the Social Sciences, ArXiv 2017.
3. Chander, A. et. al.: Working with Beliefs: AI Transparency in the Enterprise. Explainable Smart Systems Workshop, Intelligent user Interfaces, 2018.
4. Ras, G., Gerven, M., Haselager, P.: Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges, ArXiv 2018.
5. Zhang, J., Bareinboim, E. : Fairness in Decision-Making- The Causal Explanation Formula, AAAI 2018.
6. Zupan, B. et. al.: Machine learning by function decomposition. ICML 1997
7. Yang, H., Rudin, C., Seltzer, M. : Scalable Bayesian Rule Lists, ICML 2017.
8. Breiman, L. : Random Forests, Machine Learning, Volume 45, Issue 1, pp 5?32, 2001.
9. Pearl, J. : Causality: Models, Reasoning, and Inference. New York: Cambridge University Press, 2000.
10. Selvaraju, R. et.al. :Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, In International Conference on Computer Vision, 2017.
11. Son, T.: Unsupervised Neural-Symbolic Integration, XAI workshop, IJCAI 2017.
12. Hendricks, L. et. al. : Generating Visual Explanations, ECCV 2016.
13. Park, D. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence, CoRRabs//1802.08129 2018.
14. Chandrashekaran, A. et. al. : It Takes Two to Tango: Towards Theory of AI's Mind, ArXiv 2017.
15. Koh, P., Liang, P. : Understanding Black-box Predictions via Influence Functions, ICML 2017.
16. Melis, D., Jaakkola, T. : A causal framework for explaining the predictions of black-box sequence-to-sequence models, ArXiv Report 2017.
17. Bussone, A. Stumph, S, O'Sullivan, D. : The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems, IEEE Conference on Healthcare Informatics, 2015.
18. Herman, B. : The Promise and Peril of Human Evaluation for Model Interpretability, NIPS Workshop, 2017.
19. Doshi-Veklez, F., Kortz, M. : Accountability of AI Under the Law: The Role of Explanation, ArXiv 2017.
20. Narayanan, M. et. al. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. ArXiv 2018.

21. Amershi, S. et.al.: Power to the People: The Role of Humans in Interactive Machine Learning, AI Magazine, 2017.
22. Lipton, Z.: The Mythos of Model Interpretability, In International Conference on Machine Learning Workshops, 2016.
23. Millers, T., Howe, P., Sonnenberg, L.: Explainable AI: Beware of Inmates Running the Asylum, In ArXiv Report, 2017.
24. Velez, F., Kim, B.: Towards a Rigorous Science of Interpretable Machine Learning, In ArXiv Report, 2017.
25. Lombrozo, T.: The structure and function of explanations, Trends in Cognitive Science, vol 10, No. 10, 2006.
26. Rosemary, R. : What makes an explanation a good explanation? : adult learners' criteria for acceptance of a good explanation, Masters Thesis, University of Newfoundland, 1999.
27. Molineaux, M., Dannenhauer. D., Aha, D. : Towards explainable NPCs: A relational exploration learning agent. In J.C. Osborn (Ed.) Knowledge Extraction from Games: Papers from the AAAI Workshop (Technical Report WS-18-10). New Orleans, LA: AAAI Press, 2018.
28. Ribeiro, M., Singh, S., Guestrin, C. : Why Should I Trust You?": Explaining the Predictions of Any Classifier, KDD 2016.
29. Langley, P. : Explainable Agency for Intelligent Autonomous Systems., AAAI 2017.
30. Sifa, R.: Interpretable Matrix Factorization with Stochasticity Constrained Nonnegative DEDICOM, 2017.
31. Tamagnini, P.: Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations, Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics, 2017.
32. Bibal, A., Freney, B. : Interpretability of machine learning models and representations: An introduction, ESANN 2016.
33. Alonso, J.: An Exploratory Study on the Benefits of using Natural Language for Explaining Fuzzy Rule-based Systems, IEEE International Conference on Fuzzy Systems, 2017.
34. Doran, D., Schulz, S, Besold, T.: What Does Explainable AI Really Mean? A New Conceptualization of Perspectives, ArXiv 2017.
35. Brinkrolf, J. et. al. : Interpretable machine learning with reject option, at-Automatisierungstechnik 66 (4), 2018.
36. Melnikov, A.: Towards dynamic interaction-based model, ArXiv, 2018.
37. Harradon, M., Druce, J., Ruttenberg, B.: Causal Learning and Explanation of Deep Neural Networks via Autoencoded Activations. ArXiv 2018.
38. Holzinger, K.L : Can we Trust Machine Learning Results? Artificial Intelligence in Safety-Critical Decision Support, ERCIM News, 2018.
39. Alonso, J. M., Magdalena, L., Gonzalez-Rodriguez, G.: Looking for a Fuzzy System Interpretability Index: An Experimental Approach, International Journal of Approximate Reasoning, 2009.
40. Alonso, J. M. Castiello, C., Mencar, C.: A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field, International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, 2018.
41. Gacto, M.J., Alcala, R., Herrera, F.: Interpretability of Fuzzy Rule-based Systems: An Overview of Interpretability Measures, Information Sciences, 2011.
42. Dua, D., Efi, K.D.: UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, 2017.

43. Bohanec, M., Rajkovic, V.: Knowledge Acquisition and Explanation for Multi-Attribute Decision Making, International Workshop on Expert Systems and Applications, 1988.