An Empirical Investigation of Users' Assessment of XAI Explanations: Identifying the Sweet Spot of Explanation Complexity and Value

Felix Liedeker[1,*], Christoph Düsing[1], Marcel Nieveler[1] and Philipp Cimiano[1]

[1]*Semantic Computing Group, CITEC, Bielefeld University, Inspiration 1, 33619 Bielefeld, Germany*

Abstract

While the importance of explainable artificial intelligence in high-stakes decision-making is widely recognized in existing literature, empirical studies assessing users' perceived value of explanations are scarce. In this paper, we aim to address this shortcoming by conducting an empirical study focused on measuring the perceived value of the following types of explanations: plain explanations based on feature attribution, counterfactual explanations and complex counterfactual explanations. We measure an explanation's value using five dimensions: perceived accuracy, understandability, plausibility, sufficiency of detail, and user satisfaction. Our findings indicate a sweet spot of explanation complexity, with both dimensional and structural complexity positively impacting the perceived value up to a certain threshold.

Keywords

Explainable AI (XAI), Explanation Complexity, Counterfactual Explanation, User Perception

1. Introduction

In recent years, we have witnessed the prevalence of Artificial Intelligence (AI) models in various tasks, including high-stakes decision-making in domains such as clinical decision support [1] and credit risk scoring [2]. Building trust in such AI systems and complying with the respective legislation necessitates the deployment of eXplainable AI (XAI) methods to gain understanding about the inner workings of AI systems and to ensure the correctness of their output [1].

While there has been increased attention devoted to the empirical evaluation of XAI methods, user studies investigating the perceived quality, value, understandability and informativeness of explanations are scarce. We address this shortcoming by conducting an empirical study, asking users to assess the perceived value of different types of explanations for a credit risk assessment model. We measure the value of an explanation using five commonly applied dimensions (perceived accuracy, understandability, plausibility, sufficiency of detail, and user satisfaction). Our subsequent analyses focus on how the complexity of the provided explanations affects their perceived value. Accordingly, we aim to answer the following research question: *How does explanation complexity affect the perceived value of XAI explanations and is there a trade-off*

Late-breaking work, Demos and Doctoral Consortium, colocated with The 2nd World Conference on eXplainable Artificial Intelligence: July 17–19, 2024, Valletta, Malta

*Corresponding author.

fliedeker@techfak.uni-bielefeld.de (F. Liedeker); cduesing@techfak.uni-bielefeld.de (C. Düsing); mnieveler@techfak.uni-bielefeld.de (M. Nieveler); cimiano@techfak.uni-bielefeld.de (P. Cimiano)

0009-0006-2556-9430 (F. Liedeker); 0000-0002-7817-9448 (C. Düsing); 0000-0002-4771-441X (P. Cimiano)

between explanation complexity and value? To do so, we build on existing literature to introduce and empirically verify two novel notions of explanation complexity: *dimensional* and *structural complexity* of explanations, both adding to the overall *explanation complexity*.

2. Related work

The opacity of black box models and the difficulties associated with verifying or understanding their output have given rise to the subfield of XAI. Despite of the wide recognition of the importance of XAI, the concept of explainability is still underspecified [3].

Offering users explanations for decisions made by AI systems that are easy to grasp, trustworthy, and usable is a central challenge of XAI research. Numerous scholars have approached this challenge from different perspectives: By investigating how humans tend to explain and what general desiderata can be found for explanations [4], as well as examining various characteristics of explanations such as the output format of an explanation [5].

Today, counterfactual explanations (CFs) are among the most popular explanation methods, since CFs resemble the way humans naturally explain decisions [6]. Various algorithms for the generation of CFs have been proposed [7]. In addition, a plethora of different metrics and algorithms have been introduced [8] to automatically evaluate explanations - though human evaluation of explanations is still considered the gold standard [9].

Previous studies investigated the influence of the generation process of explanations on user perception (cf. e.g. [10, 11]). Wang and Yin [12] uncovered that users' perception of the helpfulness of an explanation also depends on the method used for explanation generation.

Studying the effect of different generation processes can implicitly link to the investigations of explanation complexity, because different generation algorithms may have constraints on different properties of explanations such as the length or sparsity of explanations [11].

Huysmans et al. [13] empirically investigated the correlation between comprehensibility and presentation complexity of explanations. As a result, larger, i.e. more complex representations showed a decrease in answer accuracy and confidence of the participants.

3. Methods

3.1. Types of Explanations

Our focus is on the perception of explanations that differ in terms of their overall explanation complexity. For this purpose, we introduce three different types of explanations in the following:

Plain. Plain explanations - inspired by existing feature importance approaches (e.g., LIME [14]) - are the most simple type included in our study. They are created by taking the n most relevant features, to construct an explanation of the form *Because X, the prediction is P*, where X is the set of the n most relevant features.

Counterfactual Explanations. CFs are among the most popular explanation methods for XAI and the second type of explanation included in our study. They provide explanations of the form *If X had been different, the prediction would have changed from P to Q*. Thus, they explain a decision indirectly by providing a hypothetical, but similar counterexample [15].

Complex Counterfactual Explanations. Inspired by the recently growing interest in semifactual explanations (i.e., changes that do not change the output decision [16]), we define complex CFs (CCFs) as explanations of the form *Even if X would be different, but Y would be different, the prediction is still P.* By adding an *Even if*-clause to these explanations, we deliberately increase their complexity, allowing us to test for the impact of complexity later on.

3.2. Explanation Complexity

In the realm of task complexity, it is differentiated between *presentation complexity* [17] and *domain complexity* [18]. *Presentation complexity* is induced by different information presentation formats, whereas context and environment as well as the dimensionality of data [19] contribute to the domain complexity. Since we used a single data set in our study, the domain complexity is fixed. We expand the notion of *presentation complexity* to *dimensional* and *structural complexity* to allow fine-grained control of *explanation complexity* in our study. In the following, we motivate our choice of these notions of complexity, describe which explanation features affect them, and how they contribute to overall *explanation complexity*:

Dimensional Complexity. The presumption that shorter explanations are more comprehensible is widely established in the field of XAI [4]. The rationale here is that explanations containing fewer features (i.e. shorter explanations) require less cognitive effort from users to be understood easily [4]. Our decision to define *dimensional complexity* as a contributing factor to *explanation complexity* is motivated by the assumption that requiring greater cognitive effort indicates greater *explanation complexity*. For the sake of this work, we measure *dimensional complexity* as the number of features contained in each of the described types of explanations. Hence, adding further features to an explanation implies an increased *dimensional complexity*.

Structural Complexity. In addition to the length of explanations, its type and method of generation are known to affect the explanation complexity, too [19]. In this vein, we define *structural complexity* as a function of the type of explanation provided. Here, *structural* refers to the structure of the explanation that is presented to the user and does not account for the complexity of the explanation generation method itself. Previous studies found that users can comprehend *plain* explanations more easily and tend to trust them more [12]. Consequently, we assign *CFs* a higher *structural complexity* than *plain* explanations. With respect to *CCFs*, we previously claimed that we deliberately designed them to be more complex than *CFs*. Thus, we argue that their *structural complexity* is the highest among the three types.

4. User Study

4.1. Data and Model

Our study is based on a binary classification problem using the German Credit Data [20]. The data set contains credit and customer information as well as the credit risk. Given the complexity of the original data set, we use a simplified version of it[1] and process it further by dropping samples with missing values and the *Checking Account* feature. *Duration* and *Credit amount* are encoded as categorical instead of numeric values. Amounts were originally stated in *Deutsche*

[1] https://www.kaggle.com/datasets/uciml/german-credit, last accessed: 10.01.2024

Mark, but were changed to *United States Dollars* (USD) for the ease of the English-speaking study participants[2]. The processed data set contains 522 samples and is used to train a simple neural network that achieves 67.38% accuracy. It is noteworthy that accuracy is not that important in our setting since we are interested in the explanation quality instead.

4.2. Study Design

The different explanation types described in Section 3 are evaluated by the participants in our study. Explanations are generated in the following fashion: **Plain**: Calculate the most relevant features with LIME [14] and pick the n most important features. **CFs**: Calculate CFs for all samples with the open-source Python library *Alibi Explain* [21]. Filtering is then applied to only include CFs that are feasible, e.g. sex does not change or age does not decrease in the CF instance. **Complex CFs**: A combination of both previous explanations is handcrafted.

We use the following data for the study: 9 different explanations (our three types of explanations, each with one, two or three features included in the explanation) are generated for 12 samples from the entire data set. Each participant is randomly assigned an explanation type and rates 8 different explanations with differing numbers of features of this explanation type.

Participants are asked to rate each explanation on the following questions (with the respective quality dimension in bold). Questions are answered on a 4-point Likert scale, from 1 (*Definitely YES*) to 4 (*Definitely NOT*) plus *I don't know*.

- **Perceived Accuracy**: Is the class predicted by the model accurate?
- **Understandability**: Is the provided explanation understandable?
- **Plausibility**: Is the provided explanation plausible?
- **Sufficiency of Detail**: Has the provided explanation sufficient detail?
- **User Satisfaction**: Is the provided explanation satisfying?

Our online study was conducted in early 2024 using *Prolific.com* and was designed such that participation should take around 8 minutes (Prolific reported a median completion time of 7 min and 37 s). Participants received a monetary reward of £1.50[3]. Overall, 280 participants took part in the study. 166 (60.81%) were male, 102 (37.36%) female and 5 non-binary (1.83%). The age of the participants ranged from 18 to 77 years (M = 34.20, SD = 12.17). 61.17% of participants are higher education degree holders. When asked whether the task was difficult, 67.77% of participants answered *Rather NOT* or *Definitely NOT*. Participants reported a mean experience with ML of 2.92 (SD = 0.8) on a scale from 1 (No experience) to 4 (Extensive experience).

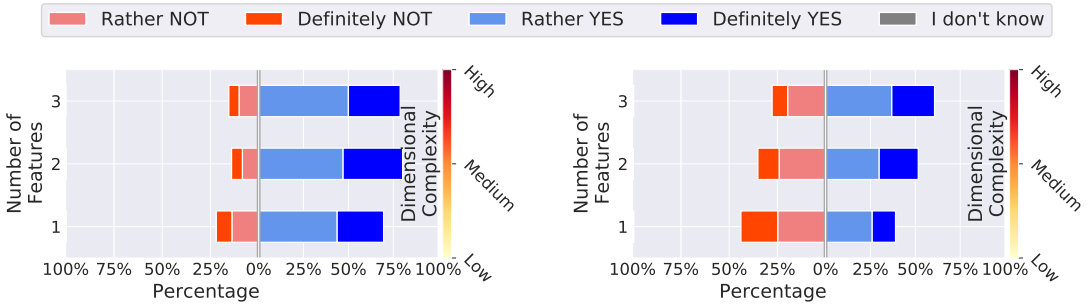
4.3. Preliminary Results

In the following, we will present the preliminary results from the analyses we performed on our study results. We measure the perceived value of explanations in our study using the five dimensions *perceived accuracy*, *understandability*, *plausibility*, *sufficiency of detail*, and *satisfaction*. In the following, we focus in particular on understandability as well as sufficiency of

[2]As a matter of fact, amounts in *Deutsche Mark* in 1994, adjusted for inflation, are almost equal to *USD* today.
[3]To match the current German minimum wage of €12.41 per hour.

detail and highlight the results for these dimensions specifically. We chose these two dimensions as they illustrate the effects of the different notions of complexity on the perceived value very well. Moreover, we observe very similar patterns for the remaining dimensions.

On Dimensional Complexity. In Figure 1, we provide the user feedback regarding the understandability and sufficiency of detail for explanations. The plot shows the percentage of participants for each of the possible answers. Here, we group explanations according to the number of features contained in it which ultimately determines their *dimensional complexity*.



(a) Understandability

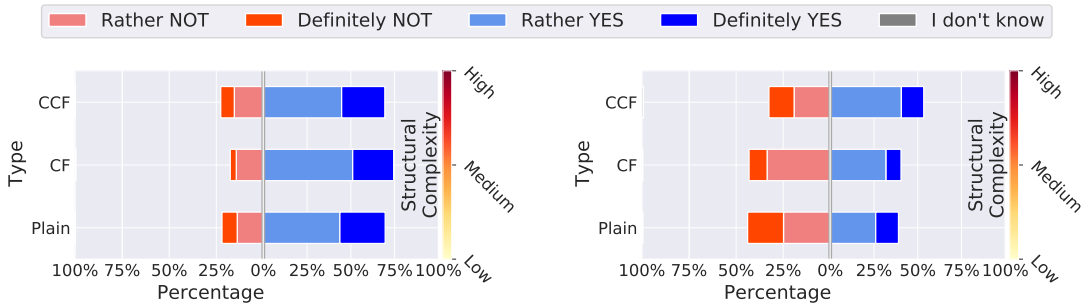
(b) Sufficiency of Detail

Figure 1: Dimensional Complexity on Explanation Value

The results in terms of understandability in Figure 1 (a) show that despite the increased *dimensional complexity*, explanations containing more than one feature are perceived as more understandable (>80% answered *Rather YES* or *Definitely YES*) than those containing a single feature only (<75% positive answers). Additionally, explanations with three features are slightly less understandable than those with two features, indicating a sweet spot at two features. Figure 1 (b), on the other hand, shows that participants considered significantly fewer explanations to be sufficiently detailed: Only about 35% for explanation with one feature, 50% with two features, and 60% with three. Furthermore, they favor explanations of higher *dimensional complexity* w.r.t. the sufficiency of detail. In contrast, explanations containing three features are also significantly better evaluated than those with two features only. Accordingly, the overall value and the perceived sufficiency of detail in particular increases for explanations of medium or high *dimensional complexity*. Explanations mentioning a single feature only have the lowest perceived value among all dimensions. The Kruskal-Wallis test finds a significant difference between explanations of different dimensional complexity for understandability ($H=11.11$, $p<0.01$) as well as the sufficiency of detail ($H=27.19$, $p<0.01$). While these findings seem to contradict existing works that assumed explanations of low sparsity to be best (e.g., [1]), they are in line with Keane and Smyth [22], who argued that explanations of moderate sparsity allow humans to get a better grasp on the concepts underlying the decision process.

On Structural Complexity. Next, we investigate how *structural complexity* correlates with explanation value. Figure 2 contains participants' feedback aggregated by the type of explanation. We sort them according to their *structural complexity* as explained in Section 3.2.

Figure 2 (a) shows that both *plain* explanations and *CCFs* are understandable for about 70% of the participants. For *CFs*, even 75% of the users agree on the fact that explanations are



(a) Understandability

(b) Sufficiency of Detail

Figure 2: Structural Complexity on Explanation Value

understandable, making explanations of medium *structural complexity* most understandable. Again, the findings are statistically significant for understandability ($H=44.92$, $p<0.01$) and sufficiency of detail ($H=10.15$, $p<0.01$). Regarding the perceived sufficiency of detail, we observe that *CCFs* are considered sufficiently detailed by more than half of the participants. For *plain* and *CFs*, only about 40% of participants agree that explanations are of sufficient detail.

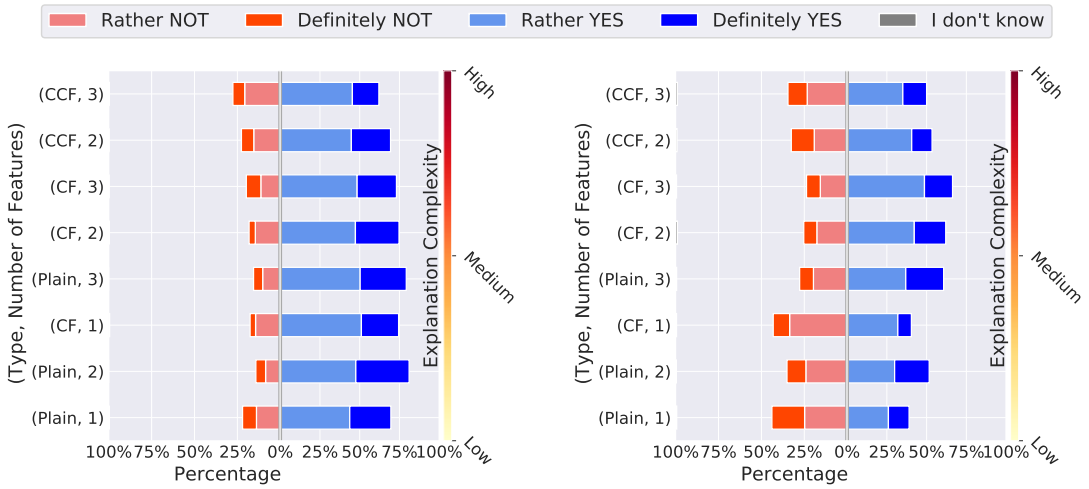
On Explanation Complexity. Finally, we study the impact of overall *explanation complexity* on the perceived value of explanations by grouping the feedback for each combination of explanation type and number of features and sort them by their *explanation complexity*.

From Figure 3 (a) we learn that explanations of particularly low and high complexity are the least understandable (<70% agreement). It furthermore shows that explanations with low to medium *explanation complexity* are most understandable (about 75% agreement). Additional interesting findings are: (1) *plain* explanations become strictly more understandable with an increasing number of features, (2) *CFs* receive similar evaluations, regardless of the number of features, and (3) *CCFs* become less understandable when adding more features.

In terms of the sufficiency of detail, Figure 3 (b) confirms previous findings on explanations with lowest and highest complexity. They are also among those considered least detailed (<50% agreement). In contrast to the findings on understandability, however, we find explanations with medium to high complexity are considered best when it comes to the sufficiency of detail.

5. Conclusion and Outlook

In this paper, we investigated users' perception of the value of different types of explanations and focused in particular on the correlation with explanation complexity. For this purpose, we conducted a user study to collect the feedback from 280 participants. Our preliminary findings suggest that both *dimensional* and *structural complexity* correlate positively with the value of explanations. In particular, users perceive explanations of either high *dimensional* or *structural complexity* as more detailed and explanations with medium to high *dimensional* or *structural complexity* as better understandable. Finally, our findings regarding the overall complexity identify a sweet spot of *explanation complexity* at medium complexity. Such explanations receive the overall best evaluation by our study participants in all dimensions of an explanation's value.



(a) Understandability

(b) Sufficiency of Detail

Figure 3: Explanation Complexity on Explanation Value

However, we acknowledge that our current findings are limited to a single task and only three different types of explanations. Therefore, in future work, we intend to extend our previous user study to contain additional use cases, tasks, and types of explanations. This will allow us to also take the *domain complexity* into account. We also aim to measure the alignment of the perceived explanation value with metrics of explanation quality commonly used in automated explanation evaluation. Finally, we seek to formalize our empirical findings and propose a taxonomy of *explanation complexity* that contributes to effective explanation design.

Acknowledgments

This work was partially funded by the *Deutsche Forschungsgemeinschaft*: TRR 318/1 2021 – 438445824 and the *Ministry of Culture and Science of North Rhine-Westphalia*: NW21-059A SAIL.

References

[1] C. Düsing, P. Cimiano, Federated learning to improve counterfactual explanations for sepsis treatment prediction, in: International Conference on AI in Medicine, Springer, 2023, pp. 86–96.

[2] J. P. Noriega, L. A. Rivera, J. A. Herrera, Machine Learning for Credit Risk Prediction: A Systematic Literature Review, Data 8 (2023) 169.

[3] T. Speith, A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods, Conference on Fairness, Accountability, and Transparency (2022) 2239–2250.

[4] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.

[5] G. Vilone, L. Longo, Classification of Explainable Artificial Intelligence Methods through Their Output Formats, Machine Learning and Knowledge Extraction 3 (2021) 615–661.
[6] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.
[7] I. Stepin, J. M. Alonso, A. Catala, M. Pereira-Fariña, A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence, IEEE Access 9 (2021) 11974–12001.
[8] V. Singh, K. Cyras, R. Inam, Explainability metrics and properties for counterfactual explanation methods, in: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, Springer, 2022, pp. 155–172.
[9] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608 (2017).
[10] J. Aechtner, L. Cabrera, D. Katwal, P. Onghena, D. P. Valenzuela, A. Wilbik, Comparing User Perception of Explanations Developed with XAI Methods, in: IEEE International Conference on Fuzzy Systems, 2022, pp. 1–7.
[11] M. Förster, P. Hühn, M. Klier, K. Kluge, User-centric explainable AI: Design and evaluation of an approach to generate coherent counterfactual explanations for structured data, Journal of Decision Systems 32 (2023) 700–731.
[12] X. Wang, M. Yin, Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making, in: International Conference on Intelligent User Interfaces, 2021, pp. 318–328.
[13] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, B. Baesens, An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models, Decision Support Systems 51 (2011) 141–154.
[14] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.
[15] Y. Jia, J. McDermid, I. Habli, Enhancing the value of counterfactual explanations for deep learning, in: International Conference on AI in Medicine, Springer, 2021, pp. 389–394.
[16] E. Kenny, W. Huang, The utility of "even if" semifactual explanation to optimise positive outcomes, Advances in Neural Information Processing Systems 36 (2024).
[17] C. Speier, The influence of information presentation formats on complex task decision-making performance, International Journal of Human-Computer Studies (2006) 1115–1131.
[18] J. Swait, W. Adamowicz, The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Strategy Switching, Journal of Consumer Research 28 (2001) 135–148.
[19] L. Weber, S. Lapuschkin, A. Binder, W. Samek, Beyond explaining: Opportunities and challenges of xai-based model improvement, Information Fusion 92 (2023) 154–176.
[20] H. Hofmann, Statlog (German Credit Data), UCI Machine Learning Repository, 1994.
[21] J. Klaise, A. V. Looveren, G. Vacanti, A. Coca, Alibi explain: Algorithms for explaining machine learning models, Journal of Machine Learning Research 22 (2021) 1–7.
[22] M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), in: International Conferenc of Case-Based Reasoning Research and Development, Springer, 2020, pp. 163–178.