PRISM: A Pragmatic Framework for Evaluating Counterfactual Explanations in XAI

Leila Methnani[1], **Virginia Dignum**[1], **Andreas Theodorou**[2]
[1]Umeå University
[2]Universitat Politècnica de Catalunya
leila.methnani@umu.se, virginia@cs.umu.se, andreas.theodorou@upc.edu

Abstract

This paper introduces PRISM, a framework for evaluating counterfactual explanations in Explainable Artificial Intelligence (XAI) through the lens of conversational pragmatics. By mapping evaluation metrics to Grice's cooperative principles—quantity, quality, relation, and manner—PRISM highlights how explanation violations can encourage users to infer deeper meanings. Demonstrated through a dashboard applied to counterfactuals from income prediction data, PRISM emphasises the social and interactive aspects of AI explanations, fostering iterative understanding and advancing human-centric XAI. Currently, we are refining the approach through qualitative studies with real-world applications.

Introduction

The societal impacts of Artificial Intelligence (AI) together with its lack of interpretability are increasingly being recognised as a concern, positioning eXplainable AI (XAI) as an important area of research (Barredo Arrieta et al. 2020). Furthermore, *counterfactual* XAI methods are gaining traction in the literature because they are easily digestible and practically useful—an important consideration for advancing *interpretability* together with *contestability* (Wachter, Mittelstadt, and Russell 2017). However, as it stands today, there is no standardised approach for evaluating counterfactual explainers. Instead, several properties are often considered and measured to indicate good performance, though the choice of properties to optimise may vary significantly (Guidotti 2022). Furthermore, the social aspect of explanations still warrants more attention (Miller 2019); fostering interactivity in XAI promotes human centricity by facilitating a dialogue that may lead to deeper insights through iterative exploration.

In this demo, we present the PRISM framework, which addresses these gaps by 1) framing an explanation as part of a dialogue and 2) showing that this framing benefits from the mapping of evaluation metrics to Grice's four maxims of conversation: quantity, quality, relation, and manner (Grice 1975). Grice's theory states that speakers engaged in conversation typically abide by these maxims; violating one or more results in a search for meaning beyond what is uttered.

In fact, the linguistic study of *pragmatics* considers how "utterances have meanings in situations" (Leech 2016) and highlights the role of inferences in communication. Using Pragmatics, Inferences, and Subtext analysis through Maxims (PRISM), our objective is to showcase a potential shift from traditional evaluation of performance to an approach that seeks to understand if deviations from metrics, when mapped to maxims, still carry meaningful insights.

PRISM Framework

In order to harness the effects of Grice's theory, we begin with a mapping of common evaluation metrics to the maxims of conversation, as shown in Table 1. We then propose making PRISM accessible through the development of an interactive dashboard depicted by Figure 1. We illustrate the use of our framework on counterfactual explanations generated for the classification task of income prediction on the Adult dataset.

Table 1: Mapping Evaluation Metrics to Gricean Maxims of Conversation

Metric	Quantity	Quality	Relation	Manner
Validity			✓	
Proximity				✓
Sparsity	✓			
Feasibility		✓		
Actionability		✓		
Diversity	✓			
Efficiency				✓
Stability				✓

Implementation Details

Our implementation, done in Python using Scikit-learn and DiCE ML, included: training a random forest classifier on the Adult income dataset, fitting the DiCE explainer model using randomized sampling, evaluating explanations against metrics for sparsity, feasibility, validity, and proximity, and interpreting scores for possible implicatures.

Dataset We used a version of the *UCI Adult dataset*, originally comprising over 48,000 entries from the 1994 US Cen-

sus database, that was pre-processed in the DiCE ML interpretability package to include 26,000 samples with eight key features: two continuous variables (age and hours worked per week) and six categorical variables (workclass, education, marital status, occupation, race, and gender). The target variable indicates income levels, with 0 representing $\leq \$50K$ and 1 representing $> \$50K$. Notably, the dataset is imbalanced, with only 24% of samples falling into the $> \$50K$ category, which could introduce bias favouring the majority class during model training and evaluation.

Model A random forest classifier was trained on the dataset, and DiCE generated counterfactuals using two model-agnostic approaches: randomized sampling and genetic algorithms. DiCE ensures *diversity* and *feasibility* while addressing proximity and sparsity in counterfactual examples. The approach has been implemented and is maintained as a Python package, which we use for our practical demonstration [1].

Dashboard The dashboard depicted in 1 and built using Plotly Dash, displays queries and counterfactuals interactively. Users can generate up to four counterfactual examples per query, evaluated across Gricean maxims and displayed via radar plots. The DiCE generated examples are then measured primarily for *sparsity*, *feasibility*, *validity*, and *proximity*, which were the four subsets of each maxim.

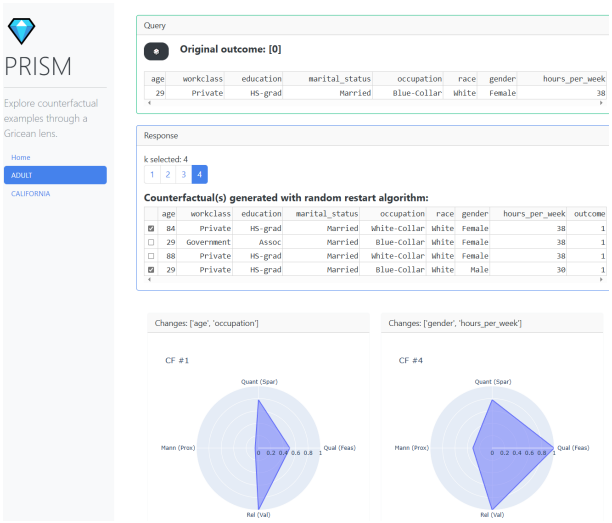

Figure 1: PRISM Screenshot showing Adult dataset

Example Figure 1 depicts a side-by-side comparison of examples that alter income predictions from from $\leq 50K$ to $> 50K$. One example adjusted 'gender' and reduced 'hours per week,' improving feasibility and proximity. Another altered 'occupation' and significantly increased 'age', resulting in lower feasibility and proximity. These variations highlight how an example, while not actionable because of alterations of the *gender* attribute, may still carry meaning within the PRISM framework.

[1]DiCE documentation available at: https://interpret.ml/DiCE/

Figure 2: Example evaluation scores.

Figure 2 illustrates the PRISM scores of explanations generated to alter income predictions in both directions, from $\leq 50K$ to $> 50K$ (blue trends) and vice-versa (red trends), over increasing k number of counterfactuals. Notably, only the blue trend showed consistently high performance in *relation*, while the red trend failed to produce relevant examples in some cases.

Conclusion and Future Work

We demonstrated through PRISM, how the study of pragmatics within linguistics can inform the ways in which humans seek meaning in information exchanged through AI explainability. We proposed re-framing evaluation as a dialogue as a stepping stone towards interactive XAI. The continuation of this work will use insights from qualitative evaluations of the platform with human subjects. Understanding how users interpret these counterfactual examples through PRISM will provide the necessary insights to propel this work forward.

References

Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.

Grice, H. 1975. Logic and Conversation. *Syntax and Semantics*, 3: 43–58.

Guidotti, R. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 1–55.

Leech, G. N. 2016. *Principles of pragmatics*. Routledge.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.

Acknowledgments

LM was funded by the Sweden's Innovation Agency (Vinnova) under the project EXPLAIN (2021-04336).