Explaining predictions with enthymematic counterfactuals

Alexander Berman[1,*], Ellen Breitholtz[2], Christine Howes[3] and Jean-Philippe Bernardy[4]

Department of Philosophy, Linguistics and Theory of Science, University of Gothenburg, Sweden

Abstract

When people are subject to high-stakes decisions informed by computer models, they have a legitimate interest in understanding the basis for the model's judgements and whether actions can be taken to turn a dispreferred decision into a preferred one. For example, if an application for a loan is denied by the model, the applicant has an interest in understanding the conditions that would yield an approval. In this paper, we argue that these kinds of counterfactual (or contrastive) explanations rest on domain-specific and commonsensical principles that can be negotiated, and sketch a method for incorporating such principles in an explanatory dialogue system using enthymematic reasoning.

Keywords

explainable AI, counterfactual explanations, interactive XAI, dialogue systems, topoi, enhtymemes

1. Introduction

Consider a conversation between a bank clerk and a customer. The clerk says: "Your loan application has been declined." The customer asks "Why?" Now, how should the clerk respond? There are several options, for example describing the bank's policy for credit assessment, or highlighting some specific aspect of the customer's status that contributed to the decision.

When people ask for an explanation for an event, they are often interested in knowing why some *other* event did not occur instead [1]. Thus, the clerk might reply: "Had you earned more than €2500, we would have granted you the loan." This human inclination for *counterfactual explanations* (CEs) has informed explainable artificial intelligence (XAI), in how to best explain predictions from machine-learning (ML) models [2]. ML is becoming widely used in a number of fields including banking [3] and healthcare [4]. In such cases, people have an interest in understanding the basis for a model's judgements, and knowing what actions they can take to turn a dispreferred outcome into a preferred one. This paper discusses two particular aspects of CEs in the context of XAI: the *dialogical* nature of explanations, and the *rhetorical* elements that underpin CEs. More specifically, we present a direction of research aiming to formalise (and later implement and evaluate) explanatory dialogue systems based on enthymematic reasoning (see

1st Workshop on Bias, Ethical AI, Explainability and the role of Logic and Logic Programming, BEWARE-22, co-located with AIxIA 2022, University of Udine, Udine, Italy, 2022

*Corresponding author.

alexander.berman@gu.se (A. Berman); ellen.breitholtz@ling.gu.se (E. Breitholtz); christine.howes@gu.se (C. Howes); jean-philippe.bernardy@gu.se (J. Bernardy)

section 2) [5]. This research demonstrates how interaction between humans and ML systems can be designed to improve transparency, facilitate sense-making and empower users.

2. Related work and background

Most existing methods for generating CEs have been designed for single-shot (i.e. non-interactive) settings [6, 7]. As for interactive approaches, Sokol et al. [8] describe a dialogue system which can generate CEs for decision trees and supports why-questions with constraints such as "Disregarding my income and employment type, what can I do to get the loan?" By contrast, our approach is model-agnostic and supports negotiable feasibility assumptions.

Reasoning in dialogue involves non-logical common-sense inferences, in rhetoric referred to as *enthymemes* [5]. Enthymemes are arguments which appeal to what is in the listener's mind, and these have been linked to why-questions [9]. Enthymemes are not strictly logical and tend to be *action oriented*, that is given a certain premise (and some additional information "in the head" of the addressee) the conclusion of the enthymeme is a particular course of action. The information "in the mind" of addressees which underpins enthymematic arguments is often called *topoi* (sg. topos) – rules of thumb according to which it is acceptable to reason. Topoi underpin virtually all kinds of inferences in discourse and dialogue.

Consider for example, two colleagues walking to work, when one of them says "Let's walk along Walnut Street. It's shorter." [10]. At first glance we may not recognise this utterance as involving inference. However, if we don't assume the topos "if you are going somewhere you want to get there as fast as possible" and the topos "if a route is shorter it is faster", the utterance hardly seems meaningful, let alone valid. Rules of thumb like these are principles that we live by, and they allow us to communicate efficiently as not everything needs to be made explicit.

However, topoi do not necessarily apply to all situations – on a sunny day we might want to take the scenic route, and the shortest route is not always the fastest. In interaction we can easily reject and question topoi, and a successful answer to a why-question regarding a suggestion should tap into a topos which is acceptable to the addressee in the context.

An enthymeme might serve to persuade or mislead a listener, but the same mechanism can also make it easier for a conversational participant to accept an honest and constructive proposal made by another agent. This is particularly pertinent in the context of XAI, since the opacity of many ML models makes it impossible to communicate the causes for a prediction in a way that is both logically complete and understandable for humans. Section 3 exemplifies how enthymematic explanations of model predictions can point the user's attention in a constructive direction, while at the same time allowing the user to reject particular topoi.

3. Enthymematic counterfactuals

To manage interactive counterfactual reasoning between a user and a system, we propose a method for selecting a relevant CE in a given situation, as well as strategies for enthymematic reasoning.

We illustrate the collaborative reasoning that our proposed method supports using the

following hypothetical interaction between a user (u) and dialogue system (s):[1]

s.1: Your loan application has been declined.
u.1: Why?
s.2: Had you earned more than €2500 and had a more qualified job, it would be accepted.
u.2: Well, I probably can't get a more qualified job very easily.
s.3: Ok. The application would also be accepted if you earned more than €3000, with your current type of job.
u.3: Hmm. That is also out of reach, unfortunately.
s.4: I see. There is also age. If you were 5 years younger, we would grant you the loan.
u.4: Actually, that's really useful to know! Perhaps my partner would be eligible.

The example assumes a predictive model with three independent variables: income, age and job status. Initially, the system assumes that reducing age is not feasible, modelled as a topos. Consequently, in s.2, the system presents a CE that does not involve changing age. In u.2, the user implicitly rejects the system's example by asserting a constraint concerning job status. Subsequently, in s.3, the CE presented by the system is filtered by both age and job status. In u.3, the user conveys that changing income is also not feasible, thus imposing a new constraint. This time, the filtering based on all three variables disqualifies all generated CEs. This triggers a fallback strategy: to make the topos (in this case concerning age) explicit, and to communicate a CE conditioned on inhibiting the topos (s.4). In this example, suppressing the age-related topos happens to be relevant to the user.

3.1. CE selection

The technique for selecting the most relevant CE rests on the following assumptions: 1) a predictive model f which maps situations described by an input vector x to a target label y and 2) a generator g that samples *statistically relevant* and *similar* counterexamples, probabilistically. For example, given a loan application represented by a vector x describing the applicant's income, age and job status, f returns `accept` or `decline`; if f returns `decline` then we should generate CEs for $y =$ `accept`.

By "statistically relevant" we mean that CEs should reflect typical characteristics in the data that f was trained on, e.g. that certain combinations of income and job states are more likely than others. An example of such a method is described by Dombrowski et al. [11].

By "similar" we mean that CEs should involve a small number of feature modifications. For example, a CE that involves modifying only income is preferred over one that involves modifying both income and job status. This desideratum is particularly important for datasets with a large number of variables, since it favours CEs that are easier to communicate and comprehend [12].

On this basis, we propose to filter the CEs generated by g by (a) *topoi* conveying domain-specific and/or commonsensical principles, and (b) *constraints* gathered from the interaction with the user. For example, the topoi may convey that generally, becoming younger or earning 5 times more are unachievable. Constraints are user-specific information harvested from earlier interaction, or other data such as a previously filled in form. In the above example, viewing an income of more than €3000 as not feasible is a constraint arising from u.3.

[1] Partly adopted from [8].

3.2. Conversational strategies and capabilities

Below we present a tentative list of dialogue strategies that can support the desired behaviour. These can be implemented as update rules in an information-state update approach [13, 14].

Counterfactual question-answering To address a user's why-question about a model-based decision, the system can reply with a CE.

Enthymematic counterfactual explanation A CE of a decision can be given in the form of an enthymeme, as illustrated by all system utterances in the dialogue example above.

Constraint integration A user's stated feasibility conditions are integrated into the information state as a constraint. A statement can be explicit as in "I can't earn more than X", or implicit/deictic as in "That is out of reach", referring to a previous system utterance.

Topos elicitation When a CE cannot be given with the current set of topoi, and disregarding a topos would enable a CE, then make the topos explicit and present the CE that it enables. This strategy underlies s.4.

4. Discussion

As described in section 3, our sketch for managing enthymematic counterfactuals assumes a generator of counterfactual candidates with two desiderata: statistical relevance (with respect to the domain as such) and similarity (with respect to the specific situation at hand). Optimizing these desiderata jointly is not trivial, since they sometimes contradict each other. For example, assuming a loan applicant with a relatively unqualified job and an income of €1500, earning 10 times more would be a small change (in the sense that only one feature needs to be modified), but statistically irrelevant (since such a high income is unusual for unqualified jobs). Since this paper primarily focuses on the dialogical reasoning process, we assume a generator with the desired properties without proposing a specific generator. Nevertheless, in order to implement and evaluate our approach in the future, a concrete generator needs to be chosen (either by selecting a previous method from the literature or by proposing a novel one).

While similarity is an important aspect of CEs [12], dialogue may offer possibilities to increase the acceptable number of feature changes. For example, a CE that involves 5 feature changes may be communicated using a stepwise information presentation strategy such as "First of all, you would need to earn more than €2500" [15]. It would be informative to study the conversational strategies that humans use in these cases.[2]

More broadly, the challenges associated with explaining predictions made by opaque models – such as deep neural networks – can be taken as a reason to simply not use such models for high-stakes decisions [17]. Nevertheless, to the extent that opaque models remain in use, explanation methods are needed. Furthermore, explaining simpler and more transparent models, such as decision trees or logistic regression, is not trivial either, particularly when targeting users who are not ML experts [18]. In fact, popular XAI techniques such as LIME [19] and SHAP [20] explain an opaque model using a simpler model which approximates the behaviour

[2]For a discussion about informing the design of XAI by collecting and analysing human dialogues, see [8, 16].

of the actual model. The approach outlined in this paper is not specifically designed for opaque models and can in principle be applied to more transparent models as well.

5. Conclusions and future work

In this paper, we have outlined a model-agnostic approach to explaining predictions using enthymematic counterfactuals. One of the main benefits is that it can direct the user's attention towards counterexamples that are actionable, while at the same time allowing the principles (topoi) that underpin the reasoning to be questioned and rejected in a dialogue. In future work, we plan to formalise the approach and implement and evaluate a prototype. Another interesting avenue of research concerns the question of how to acquire the topoi, which can be encoded by the system developers, but may be harvestable from interactions [21]. More generally, we believe that enthymemes offer a useful avenue for further research in XAI and hope that the approach outlined here can stimulate further work in this direction.

Acknowledgements

This work was supported by the Swedish Research Council (VR) grant 2014-39 for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

[1] D. J. Hilton, Conversational processes and causal explanation., Psychological Bulletin 107 (1990) 65.
[2] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.
[3] X. Dastile, T. Celik, M. Potsane, Statistical and machine learning models in credit scoring: A systematic literature survey, Applied Soft Computing 91 (2020) 106263.
[4] A. Qayyum, J. Qadir, M. Bilal, A. Al-Fuqaha, Secure and robust machine learning for healthcare: A survey, IEEE Reviews in Biomedical Engineering 14 (2021) 156–180.
[5] E. Breitholtz, Enthymemes and Topoi in Dialogue: The Use of Common Sense Reasoning in Conversation, Brill, Leiden, The Netherlands, 2020.
[6] S. Verma, J. Dickerson, K. Hines, Counterfactual explanations for machine learning: A review, 2020. URL: https://arxiv.org/abs/2010.10596.
[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] K. Sokol, P. Flach, One explanation does not fit all, KI - Künstliche Intelligenz 34 (2020) 235–250.
[9] J. Schlöder, E. Breitholtz, R. Fernández, Why?, in: Proceedings of the Workshop on the Semantics and Pragmatics of Dialogue (JerSEm), 2016, pp. 5–14.

[10] M. A. Walker, The effect of resource limits and task complexity on collaborative planning in dialogue, Artificial Intelligence 85 (1996) 181–243.
[11] A.-K. Dombrowski, J. E. Gerken, K.-R. Müller, P. Kessel, Diffeomorphic counterfactuals with generative models, 2022. URL: https://arxiv.org/abs/2206.05075.
[12] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harv. JL & Tech. 31 (2017) 841.
[13] S. Larsson, Issue-based dialogue management, Ph.D. thesis, 2002.
[14] J. Ginzburg, The interactive stance: Meaning for conversation, Oxford University Press, 2012.
[15] V. Demberg, J. D. Moore, Information presentation in spoken dialogue systems, in: 11th Conference of the EACL, 2006, pp. 65–72.
[16] A. Berman, C. Howes, "Apparently acousticness is positively correlated with neuroticism". Conversational explanations of model predictions, in: Proceedings of SemDial, 2022.
[17] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.
[18] J. Dieber, S. Kirrane, Why model why? Assessing the strengths and limitations of LIME, 2020. URL: https://arxiv.org/abs/2012.00093. doi:10.48550/ARXIV.2012.00093.
[19] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, 2016, p. 1135–1144.
[20] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).
[21] V. Maraev, E. Breitholtz, C. Howes, J.-P. Bernardy, Why should I turn left? Towards active explainability for spoken dialogue systems., in: Proceedings of the Reasoning and Interaction Conference (ReInAct 2021), ACL, Gothenburg, Sweden, 2021, pp. 58–64.