Counterfactual Explanations for eXplainable AI (XAI)

Greta Warren[1,2,*,†]

[1]School of Computer Science, University College Dublin, Dublin, Ireland
[2]Insight SFI Centre for Data Analytics, University College Dublin, Dublin, Ireland

Abstract

Counterfactual explanation has become a popular and promising method of explaining black-box AI systems and their decisions in recent years. However, a lack of rigorous psychological research means that little is known about what constitutes a 'good' counterfactual explanation, or how they facilitate user understanding of the underlying system. My doctoral research aims to examine how these sorts of explanations are understood and evaluated by users, identify desirable characteristics of counterfactual explanations, and investigate how current state-of-the-art counterfactual explanation techniques satisfy these criteria. These insights will guide the development of a novel explanation method designed to meet the psychological requirements of users. In order to address these research questions, to date I have conducted three large-scale, well-controlled user studies using materials drawn from an existing case-base. These studies have yielded novel findings about the impact of counterfactual explanation on users objective understanding and subjective judgments of an AI system. Based on these results, we have proposed an extension of a case-based counterfactual method that produces psychologically-valid explanations, which is to our knowledge, the first method designed with this specific criterion in mind.

Keywords

XAI, counterfactual, contrastive, CBR

1. Introduction

Explaining opaque AI systems and their decisions using contrastive counterfactual examples has gained considerable traction in recent years (see [1, 2] for reviews). To this end, concepts from case-based reasoning (CBR) such as Nearest Unlike Neighbours (NUNs [3]) have inspired such approaches to explanation-by-example, by providing information about how an alternative system decision could have been made, had some aspect of the input data been different [4]. For example, after rejection for a bank loan, a counterfactual explanation may inform the applicant: "had your salary been €10,000 higher, your application would have been approved". Counterfactual explanations have been proposed to appeal to important characteristics of human explanation and causal reasoning [5, 6], as well as offering potential for recourse [2]. However, although there has been a surge in the number of methods proposed for generating counterfactual explanations computationally, there is limited evidence to show that the outputs of these methods meet the psychological criteria of a 'good' explanation, while a lack of controlled user studies to evaluate their impact on user understanding and perceptions of the

ICCBR DC'22: Doctoral Consortium at ICCBR-2022, September, 2022, Nancy, France
*Corresponding author.
greta.warren@ucdconnect.ie (G. Warren)
0000-0002-3804-2287 (G. Warren)

system jeopardises their real-world utility. Furthermore, many existing studies rely on users' subjective satisfaction, trust, or fairness judgments, which may not necessarily reflect the depth of their understanding of the system's causal mechanisms [7].

My doctoral research seeks to address these issues by examining how counterfactual explanations are evaluated by human users using both objective and subjective measures, and identifying psychological desiderata of these sorts of explanations. This is achieved by conducting large-scaled, controlled user studies with materials drawn from an existing case-base. These insights will guide an analysis of existing computational methods to assess how well they meet these psychological criteria, as well as the design of a novel, psychologically-grounded case-based approach to counterfactual explanation.

2. Research Plan

2.1. Research Objectives

Counterfactual explanations have received significant attention in recent years as a means of elucidating decisions made by black box AI systems to users. Over 100 methods have been proposed to generate counterfactual explanations [1], and are commonly compared to the state of the art with reference to proximity [8], sparsity [9], and plausibility [2]. However, it is striking that so few of these methods are evaluated with respect to the primary stakeholders (i.e., end-users [1]). Moreover, these quantitative metrics are based on researchers' intuitions about what constitutes a 'good' explanation, however, it is unclear how (and if) they map to longstanding psychological and philosophical definitions of explanatory power [5, 10]. Indeed, although there is a rich body of literature surrounding human explanation [10], and counterfactual reasoning [11], relatively little is known about counterfactual explanations beyond the context of XAI and how they are understood.

The core objectives of my research are to examine how counterfactual explanations impact users' understanding and perceptions of an AI system, and identify the optimal characteristics of these explanations, in order to guide the design of a novel, user-centric counterfactual method that produces psychologically-valid explanations. Specifically, I investigate how counterfactual explanations of AI predictions improve users' objective accuracy in a prediction task, and subjective judgments of explanation satisfaction and trust in the system. In addition, I examine how focusing on certain feature-types appears to increase user accuracy, and hence, help users more readily understand the AI system. These insights will guide both the development of a counterfactual explanation method that meets users' psychological requirements, as well as shed new light on counterfactual explanation in human cognition. The key research questions I have identified are:

- What are the optimal characteristics of a counterfactual explanation from a psychological perspective?
- Which counterfactual methods produce the best explanations in terms of computational metrics (e.g., sparsity, proximity, plausibility)?
- How can a counterfactual method produce explanations that meet users' psychological criteria of explanations?

2.2. Approach / Methodology

User Studies. In order to investigate the effects of counterfactual explanations on users' understanding and evaluations of an AI system, we conducted a series of user studies designed to assess the impact of counterfactual explanation on users' task accuracy and subjective judgments. We compare these effects to those of causal explanations and a control condition (in which participants receive only descriptions of the system's decisions). Participants in the studies were presented with materials in the form of case-instances, each consisting of five features used to predict blood alcohol content: gender (male/female), weight (in kg), amount of alcohol consumed by the person (in units), duration of drinking period (in minutes), and stomach-fullness (full/empty). Users were shown the output of a simulated AI system presented as an application, designed to predict whether someone is over the legal blood alcohol content limit to drive. Materials were selected from a case-base of instances of normally-distributed values of the feature-set. In the training phase of the experiments, participants were shown examples of tabular data for different individuals, and asked to make a judgment about whether each individual was under or over the limit on each screen. After giving their response, feedback was given on the next page, along with an explanation, the content of which was dependent on the experimental condition (see Figure 1 for a sample of the material used in the counterfactual condition). Upon completing the training phase, participants began the testing phase, in which they were shown more example instances referring to individuals and again asked to judge if each individual was over or under the legal limit to drive. For each instance, participants were asked to consider a specific feature in making their prediction; for instance, "Given this person's WEIGHT, please make a judgment about their blood alcohol level." After submitting their response, no feedback or explanation was given. In addition to measuring task accuracy, participants were also asked to provide judgments of explanation satisfaction and trust, measured using the DARPA Explanation Satisfaction and Trust scales [12] respectively, allowing us to evaluate explanation quality using both objective and subjective measures, which may not necessarily correspond with one another.

Figure 1: Feedback for Incorrect Answer in the Counterfactual condition of the study

Towards a Psychologically-valid Counterfactual Method. A key result from the user studies discussed above was that users were significantly more accurate when making predic-

tions about categorical features (stomach fullness and gender) than continuous features (units, weight and drinking duration). This finding is supported by evidence from the counterfactual reasoning literature that people do not spontaneously change continuous variables when generating counterfactuals for past events [13]. In light of this, we conducted an analysis of NUNs with categorical feature differences in a number or popular UCI datasets, observing that they are exceedingly rare. Hence, we developed a variation of Keane and Smyth's [9] case-based counterfactual method, which applies post-hoc transformations to the feature differences in order to produce counterfactual explanations more intuitively understandable to end-users (see [14] for more detail).

3. Progress Summary

To date, I have conducted three large-scale, well-controlled user studies (total N = 474) which have revealed novel insights into how counterfactual and causal explanations are understood and perceived by users. While counterfactual explanations are judged as more satisfying and trustworthy than causal explanations, they appear to be only slightly more effective in improving objective performance in a prediction task. This disconnect between objective and subjective measures suggests that it is critical to examine how explanations aid user understanding rather than merely improve subjective perceptions. Furthermore, users appear to understand the impact of categorical features on the system's decision more readily than that of continuous features, a distinction that current computational methods do not account for. Findings from the first user study were presented at the *Cognitive Aspects of Knowledge Representation* workshop at *IJCAI'22* [15], with preliminary results presented at *CogSci'21*. Findings from the complete series of user studies are currently being prepared for submission to a top-tier conference.

Based on the finding that counterfactuals that change categorical features are more readily understood than those focusing on continuous features, we developed a counterfactual method that accounts for this feature-type distinction. An analysis of common UCI datasets suggests that sparse counterfactuals with categorical feature-changes are relatively rare, and so our method adapts Keane and Smyth's [9] case-based technique to transform feature-differences into categorical versions, without significant decrement to performance in terms of coverage and proximity of the counterfactuals produced. To our knowledge, this is the first counterfactual method designed to meet identified psychological requirements for explanation by users, and will be presented at *ICCBR'22* [14].

The main focus of my research at present is the design of a second series of psychological experiments examining the role of simplicity (or sparsity) in counterfactual explanation, and how it impacts user understanding and subjective judgments. In tandem, I am working on the implementation and evaluation of popular counterfactual computational methods, in order to identify those methods which are most successful (i.e. have the best average performance) in generating counterfactuals that meet given criteria of an explanation over a set of representative problems. These criteria include conventional metrics (e.g., proximity, sparsity, plausibility) as well as novel properties derived from user testing (such as whether a counterfactual makes continuous or categorical feature-changes). The final phase of my Ph.D. research will involve synthesising the insights from these two strands of work in order to develop a novel, psychologically-grounded

method for counterfactual explanation.

References

[1] M. T. Keane, E. M. Kenny, E. Delaney, B. Smyth, If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques, IJCAI-21 (2021).

[2] A. H. Karimi, B. Schölkopf, G. Barthe, I. Valera, A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects, volume 1, Association for Computing Machinery, 2020.

[3] C. Nugent, D. Doyle, P. Cunningham, Gaining insight through case-based explanation, Journal of Intelligent Information Systems 32 (2009) 267–295.

[4] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, J. Wilson, The what-if tool: Interactive probing of machine learning models, IEEE transactions on visualization and computer graphics 26 (2019) 56–65.

[5] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38. doi:10.1016/j.artint.2018.07.007.

[6] R. M. Byrne, Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning, volume 2019-Augus, 2019, pp. 6276–6282. doi:10.24963/ijcai.2019/876.

[7] Z. Buçinca, P. Lin, K. Z. Gajos, E. L. Glassman, Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems, 2020, pp. 454–464. doi:10.1145/3377325.3377498.

[8] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harvard Journal of Law & Technology 31 (2018).

[9] M. T. Keane, B. Smyth, Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), Springer, Cham, 2020, pp. 163–178.

[10] F. C. Keil, Explanation and understanding, Annual Review of Psychology 57 (2006) 227–254. doi:10.1146/annurev.psych.57.102904.190100.

[11] R. M. Byrne, Counterfactual thought, Annual review of psychology 67 (2016) 135–157.

[12] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for Explainable AI: Challenges and Prospects, Technical Report December, 2018.

[13] D. Kahneman, A. Tversky, The simulation heuristic, in: D. Kahneman, P. Slovic, A. Tversky (Eds.), Judgment Under Uncertainty: Heuristics and Biases, Cambridge University Press, New York, 1982, pp. 201–8.

[14] G. Warren, B. Smyth, M. T. Keane, "better" counterfactuals, ones people can understand: Psychologically-plausible case-based counterfactuals using categorical features for explainable ai (xai), in: To appear in ICCBR'22, 2022.

[15] G. Warren, M. T. Keane, R. M. Byrne, Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in xai, in: IJCAI-22 Workshop on Cognitive Aspects of Knowledge Representation, 2022.