

User-Driven Counterfactual Generator: A Human Centered Exploration

Isacco Beretta¹, Eleonora Cappuccio^{1,2,3} and Marta Marchiori Manerba^{1,2}

¹Computer Science Department, Università di Pisa, Italy

²KDD Laboratory, ISTI, National Research Council, Pisa, Italy

³Computer Science Department, Università degli Studi di Bari Aldo Moro, Italy

Abstract

In this paper, we critically examine the limitations of the techno-solutionist approach to explanations in the context of counterfactual generation, reaffirming interactivity as a core value in the explanation interface between the model and the user.

Keywords

Human-Centered AI, XAI, Counterfactuals, XUI


1. Introduction


Progress in the performance and efficiency of automatic decision-making systems has incentivized AI-based solutions in pervasive and impactful contexts of daily life, such as finance, healthcare, and transportation. A problematic aspect of these models is the lack of explainability. It is difficult to provide the reasons behind the automated decisions due to the complexity of the process and the large amounts of data required. In some sensitive real-world contexts, accounting for the algorithmic decision is necessary for the user to understand and contest the motivations behind the system's output. This is especially important when the outcome strongly impacts human life or has harmful consequences. Moreover, recent policies such as the GDPR (General Data Protection Regulations) [1, 2] claim the relevance of appealing in case of rejection, i.e., to request what changes users should make to be accepted in receiving a positive response, for example, to a mortgage application. In this way, users can be empowered to understand how the algorithmic decision affects them. Counterfactual approaches explain individual predictions describing what-if contrastive scenarios [3, 4]. Specifically, the explanations indicate to the user the minimum change necessary in the feature space representing them so that the output of the automatic system changes toward the desired outcome. Evaluating the quality of counterfactual explanations requires careful consideration of multiple desired qualities, which have been defined in the literature as *validity*, *sparsity*, *similarity*, *plausibility*, *discriminative power*, *actionability*, *causality*, and *diversity* [5, 6, 7]. The complexity of this range of evaluation metrics

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

✉ isacco.beretta@phd.unipi.it (I. Beretta); eleonora.cappuccio@phd.unipi.it (E. Cappuccio); marta.marchiori@phd.unipi.it (M. M. Manerba)

ORCID 0000-0002-0463-6810 (I. Beretta); 0000-0002-6105-2512 (E. Cappuccio); 0000-0003-2251-1824 (M. M. Manerba)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

presents a critical challenge in comparing the currently available algorithms and establishing benchmark procedures.

Current counterfactual generation techniques do not allow for user interaction, thus limiting the practical applicability of these methods in real-world contexts. See [8] for a more detailed analysis. Moreover, several studies have highlighted the need to focus more on the user’s point of view and needs [9, 10, 11, 12]: It is crucial to rethink new ways of interaction between the users and the XAI algorithm. Specifically, as highlighted by [13, 9, 14, 11], an explanation process has to be outlined as a continuous dialogue between the explainer and the explainee. Therefore XAI has to consider interactivity as a fundamental part of the process, for example, through the use of novel user interfaces that allows model inspection at will [15]. To address these issues, we present *User-Driven Counterfactual Generator*.

2. User-Driven Counterfactual Generator

Preliminaries. Given a classifier b that outputs the decision $y = b(x)$ for an instance x , a counterfactual explainer \mathcal{E} outputs a perturbed instance x' such that the decision of b on x' changes, i.e., $b(x') \neq y$, and such that the cost of the action $c(x, x')$ to go from x to x' is *minimal*. Minimality refers to an abstract cost function that must be specified when implementing any concrete method. Often, the choice falls on an ℓ_p norm [8]. In the vast majority of cases, c exhibits the following properties.

- *Stationarity*: $c(x, x')$ is predetermined and not updated through user feedback. Moreover, it lacks the ability for \mathcal{E} to adapt during post-deployment usage, which would allow for fine-tuning between user needs and the tool’s effectiveness.
- *Translational Invariance*: $c(x, x')$ depends solely on the distance between x and x' , i.e., $c(x, x') = c(x' - x)$. For example, increasing the salary by 300\$ is assumed to be equally difficult for an individual earning 1000\$ and one earning 3000\$.
- *Universality*: c can not depend on exogenous factors of the system. In other words, it is assumed that the same cost function is suitable for all users: in reality, individual properties not visible to the system can influence c . For example, the ease of changing jobs may depend on the types of available activities in the residential area. The cost of an action in the real world depends on individuals and multiple, often subjective, factors. Ignoring this aspect often proves to be overly restrictive.

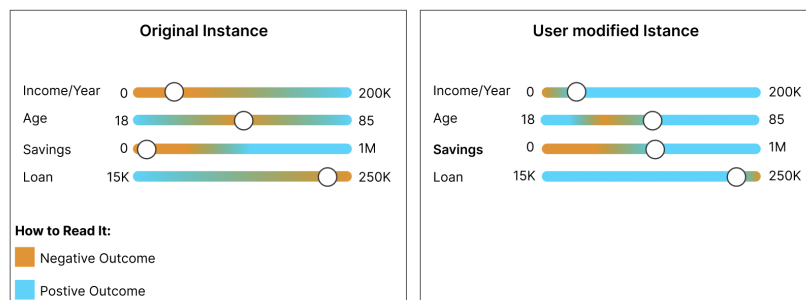


Figure 1: User’s situation requesting a loan (left); instance perturbed changing *Savings* (right).

Each of these properties imposes specific limitations on the ability to provide effective recommendations to each individual: it is precisely this gap that our contribution aims to fill.

Algorithm 1: GRID-EVALUATION(\hat{X}, x, b)

Input : \hat{X} - list of empty lists, x - user instance, b - black box
Output: \hat{Y} - list of lists containing b 's prediction on \hat{X}

```

1  $n \leftarrow \text{len}(x)$ ; // storing features number
2 for  $i = 1$  to  $n$  do
3    $n_i \leftarrow \text{len}(\hat{X}_i)$ ; // storing bins number for feature  $i$ 
4   for  $j = 0$  to  $n_i$  do
5      $\hat{x} \leftarrow \text{copy}(x)$ ; // creating a copy of  $x$ 
6      $\hat{x}_i \leftarrow j/n_i$ ; // changing  $\hat{x}$ 's  $i$ -th feature with bin value
7      $\hat{X}_{i,j} \leftarrow \hat{x}$ ; // saving the perturbed instance
8  $\hat{Y} \leftarrow b(\hat{X})$ ; // predicting scores
9 return  $\hat{Y}$ 

```

Proposed Approach. Our method proposes an alternative formulation of the counterfactual generation problem, departing from the algorithmic perspective and instead shifting the focus to user interaction. Our approach is a local, agnostic, post-hoc explanation method designed to account for any black box model. It requires access to the model’s probability outputs and is specifically tailored for tabular data with a binary target variable. Through a visual interface, users are free to explore and generate autonomously effective counterfactuals according to their needs without the mediation of an explanation algorithm between the human and the decision model, empowering users with complete autonomy and control over the process. As there is no ground truth available w.r.t. explanation, evaluating the explanations is qualitative and relies on user tests to assess their effectiveness. The interface, depicted in Figure 1, allows users to independently modify the value of one of the features that characterize the instance x . This can be done by adjusting the slider *feature by feature* within its designated range¹. Each slider is characterized by a colored gradient representing the model’s scores for the different values of the feature. Users can choose which feature to modify by moving the cursor to a value that improves their score, selecting the feature they find most comfortable to change. After each adjustment, the instance x is promptly updated with new values, generating new gradients and enabling further modifications until the user is content with the set of changes and insights. We present a graphical representation in Figure 2, i.e., the allowed movements using sliders. The user can move along each axis in the feature space but cannot make diagonal movements. The cumulative nature of the modifications ensures the ability to reach any point in the space. In Algorithm 1, we report the pseudo-code for a single step of the process². The number of predictions required scales linearly with the number of features, ensuring the interface update remains essentially real-time. The evaluation of the method’s effectiveness can

¹In the case of categorical features, the slider is replaced by a drop-down menu that presents the available discrete values for that feature. For simplicity, it is not included in Figure 1.

²For the sake of simplicity, we assume that all features are normalized and have values between 0 and 1.

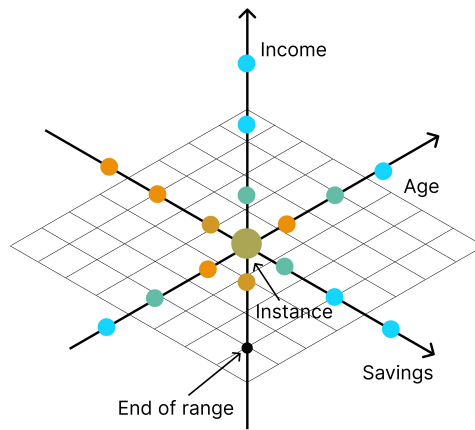


Figure 2: Allowed movements through sliders.

not rely solely on data-driven metrics mentioned in Section 1, as generating a counterfactual depends entirely on user interaction. Additionally, our approach does not assume the existence of a pre-established cost function, making a quantitative assessment of its optimality not useful. Therefore, a direct comparison with algorithmic methods should be dismissed. Embracing a human-centered perspective, the evaluation requires a conceptual change by involving users within it in order to understand whether the explanation is convincing for the intended user, the ultimate stakeholder in the process [16, 17]. Concretely, a qualitative evaluation in the form of a user study is necessary. This test will provide a measure of human validation regarding the comprehensibility and usability of the interface, as well as the usefulness of exploration as a means of generating explanations for the underlying decision model.

3. Conclusion

In this paper, we have critically examined the limitations of the techno-solutionist approach to explanations in the context of counterfactual generation, reaffirming interactivity as a core value in the interface between the model and the user. By embracing a user-centric perspective, the field of XAI can overcome the drawbacks of a purely technologically driven perspective. This approach holds the potential to not only enhance the interpretability and transparency of AI models [18] but also foster trust and effective decision-making in human-AI interactions [19].

Acknowledgments

This work has been supported by the European Community Horizon 2020 programme under the funding scheme ERC-2018-ADG G.A. 834756 *XAI: Science and technology for the eXplanation of AI decision making* and by the European Union’s Horizon Europe Programme under the CREXDATA project, grant agreement no. 101092749.

References

- [1] F. Sovrano, F. Vitali, M. Palmirani, Making things explainable vs explaining: Requirements and challenges under the GDPR, CoRR abs/2110.00758 (2021). URL: <https://arxiv.org/abs/2110.00758>. arXiv:2110.00758.
- [2] O. Seizov, A. Wulf, Artificial intelligence and transparency: A blueprint for improving the regulation of ai applications in the eu, *European Business Law Review* 31 (2020) 611–640. doi:10.54648/EULR2020024.
- [3] A. Karimi, B. Schölkopf, I. Valera, Algorithmic recourse: from counterfactual explanations to interventions, in: FAccT, ACM, 2021, pp. 353–362.
- [4] C. Molnar, G. Casalicchio, B. Bischl, Interpretable machine learning - A brief history, state-of-the-art and challenges, in: ECML PKDD 2020 Workshops - Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, volume 1323 of *Communications in Computer and Information Science*, Springer, 2020, pp. 417–431. URL: https://doi.org/10.1007/978-3-030-65965-3_28. doi:10.1007/978-3-030-65965-3_28.
- [5] Y. Jia, J. Bailey, K. Ramamohanarao, C. Leckie, M. E. Houle, Improving the quality of explanations with local embedding perturbations, in: A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, G. Karypis (Eds.), Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4–8, 2019, ACM, 2019, pp. 875–884. URL: <https://doi.org/10.1145/3292500.3330930>. doi:10.1145/3292500.3330930.
- [6] Y. Zhang, K. Song, Y. Sun, S. Tan, M. Udell, " why should you trust my explanation?" understanding uncertainty in lime explanations, arXiv preprint arXiv:1904.12991 (2019).
- [7] R. Guidotti, Evaluating local explanation methods on ground truth, *Artif. Intell.* 291 (2021) 103428. URL: <https://doi.org/10.1016/j.artint.2020.103428>. doi:10.1016/j.artint.2020.103428.
- [8] A. Karimi, G. Barthe, B. Schölkopf, I. Valera, A survey of algorithmic recourse: Contrastive explanations and consequential recommendations, *ACM Comput. Surv.* 55 (2023) 95:1–95:29. URL: <https://doi.org/10.1145/3527848>. doi:10.1145/3527848.
- [9] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, *Artif. Intell.* 267 (2019) 1–38. URL: <https://doi.org/10.1016/j.artint.2018.07.007>. doi:10.1016/j.artint.2018.07.007.
- [10] Q. V. Liao, K. R. Varshney, Human-centered explainable AI (XAI): from algorithms to user experiences, CoRR abs/2110.10790 (2021). URL: <https://arxiv.org/abs/2110.10790>. arXiv:2110.10790.
- [11] A. M. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, M. S. Kankanhalli, Trends and trajectories

for explainable, accountable and intelligible systems: An HCI research agenda, in: CHI, ACM, 2018, p. 582.

- [12] S. Amershi, M. Cakmak, W. B. Knox, T. Kulesza, Power to the people: The role of humans in interactive machine learning, *AI Mag.* 35 (2014) 105–120. URL: <https://doi.org/10.1609/aimag.v35i4.2513>. doi:10.1609/aimag.v35i4.2513.
- [13] P. Madumal, T. Miller, L. Sonenberg, F. Vetere, A grounded interaction protocol for explainable artificial intelligence, in: E. Elkind, M. Veloso, N. Agmon, M. E. Taylor (Eds.), *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS '19*, Montreal, QC, Canada, May 13-17, 2019, International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1033–1041. URL: <http://dl.acm.org/citation.cfm?id=3331801>.
- [14] M. Chromik, M. Schuessler, A taxonomy for human subject evaluation of black-box explanations in XAI, in: A. Smith-Renner, S. Kleanthous, B. Y. Lim, T. Kuflik, S. Stumpf, J. Otterbacher, A. Sarkar, C. Dugan, A. Shulner-Tal (Eds.), *Proceedings of the Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies co-located with 25th International Conference on Intelligent User Interfaces (IUI 2020)*, Cagliari, Italy, March 17, 2020, volume 2582 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: <https://ceur-ws.org/Vol-2582/paper9.pdf>.
- [15] B. Shneiderman, Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems, *ACM Trans. Interact. Intell. Syst.* 10 (2020) 26:1–26:31. URL: <https://doi.org/10.1145/3419764>. doi:10.1145/3419764.
- [16] Q. V. Liao, Y. Zhang, R. Luss, F. Doshi-Velez, A. Dhurandhar, Connecting algorithmic research and usage contexts: A perspective of contextualized evaluation for explainable AI, in: J. Hsu, M. Yin (Eds.), *Proceedings of the Tenth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2022*, virtual, November 6-10, 2022, AAAI Press, 2022, pp. 147–159. URL: <https://ojs.aaai.org/index.php/HCOMP/article/view/21995>.
- [17] Z. J. Wang, J. W. Vaughan, R. Caruana, D. H. Chau, GAM coach: Towards interactive and user-centered algorithmic recourse, in: A. Schmidt, K. Väänänen, T. Goyal, P. O. Kristensson, A. Peters, S. Mueller, J. R. Williamson, M. L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023*, Hamburg, Germany, April 23-28, 2023, ACM, 2023, pp. 835:1–835:20. URL: <https://doi.org/10.1145/3544548.3580816>. doi:10.1145/3544548.3580816.
- [18] S. Larsson, F. Heintz, Transparency in artificial intelligence, *Internet Policy Rev.* 9 (2020).
- [19] C. Nicodeme, Build confidence and acceptance of ai-based decision support systems - explainable and liable AI, in: *HSI, IEEE*, 2020, pp. 20–23.