*HHAI 2024: Hybrid Human AI Systems for the Social Good*
*F. Lorig et al. (Eds.)*

*doi:10.3233/FAIA240184*

# An Interactive Interface for Feature Space Navigation

Eleonora CAPPUCCIO [a,b,c,1], Isacco BERETTA [a], Marta MARCHIORI MANERBA [a,c]
and Salvatore RINZIVILLO [c]
[a] *Università di Pisa*
[b] *Università degli Studi di Bari Aldo Moro*
[c] *CNR-ISTI*
ORCiD ID: Eleonora Cappuccio https://orcid.org/0000-0002-6105-2512, Isacco Beretta https://orcid.org/0000-0002-0463-6810, Marta Marchiori Manerba https://orcid.org/0000-0003-2251-1824, Salvatore Rinzivillo https://orcid.org/0000-0003-4404-4147

**Abstract.** In this paper, we present `Feature Space Navigator`, an interactive interface that allows an exploration of the decision boundary of a model. The proposal aims to overcome the limitations of the techno-solutionist approach to explanations based on factual and counterfactual generation, reaffirming interactivity as a core value in designing the conversation between the model and the user. Starting from an instance, users can explore the feature space by selectively modifying the original instance, on the basis of her own knowledge and experience. The interface visually displays how model predictions react in response to the adjustments introduced by the users, letting them to identify relevant prototypes and counterfactuals. Our proposal leverages the autonomy and control of the users that can explore the behavior of the decision model accordingly with their own knowledge base, reducing the need for a dedicated explanation algorithm.

**Keywords.** Human-Centered AI, Human-AI Interaction, Interactive ML, Explainable AI, Counterfactuals, Explanation User Interface

## 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 algorithms and the large amounts of data processed. This lack of transparency and accountability can lead to mistrust and concerns about the ethical implications of using AI in decision-making [1]. 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, relevant

[1] Corresponding Author: Eleonora Cappuccio, eleonora.cappuccio@phd.unipi.it.

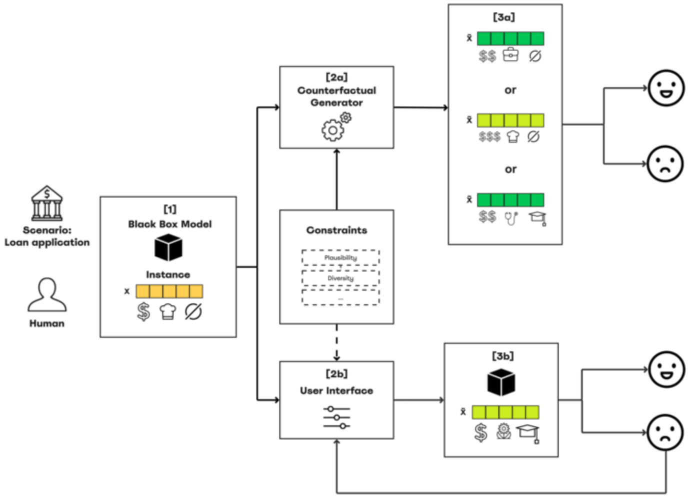


**Figure 1.** Workflow of `Feature Space Navigator`.

policies such as the GDPR (General Data Protection Regulations) [2,3] 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 loan application. In this way, users can be empowered to understand how the algorithmic decision affects them.

In this work, we present `Feature Space Navigator`, an interactive interface that allows an exploration of the decision boundary of any black box model. We report in Fig. 1 a visual workflow of our approach. Imagine a person applying for a loan to a bank that leverages an automatic decision-making system. The instance $x$ represents her current situation (stage [1] of the workflow): she has a certain amount of capital, her current working position is being a chef in a restaurant, and she has no degree. The decision emitted by the black box model is negative, hence her request is denied.

Following the diagram, in stage [2a] an automatic counterfactual technique generates instances trying to optimize a predefined cost function that embeds a set of constraints, e.g., *plausibility* and *diversity*. Three counterfactual instances receiving a positive outcome are proposed to the user (stage [3a]). The scenarios suggested are: (i) the person has a medium amount of capital and occupies a business role, (ii) the person remains a chef, but she has a high amount of capital, and (iii) the person has a medium amount of capital, she is a doctor and she has a degree in medicine. From the counterfactuals returned, the user satisfaction is symbolized by the smiley and gloomy faces. Following this pipeline, the subjective human satisfaction is impossible to capture and it remains unknown or an irrelevant feedback.

Our proposal, exemplified in stage [2b], consists of an interface that lets the users freely explore their options, changing the features at will. The set of constrains is not directly embedded nor expressed through the interface, but it is implicitly delegated to the common sense of the users. The resulting counterfactual user-crafted in stage [3b] has the same amount of money, by profession she is a florist and has a degree. The difference between our approach w.r.t. the traditional one consists in the possibility for the user to intervene again on the interface and generate a new, more effective counterfactual if the one created is not useful or applicable, creating a continuous loop through the interaction.

The rest of the paper is organized as follows. In Section 2, we report a critical literature discussion and a review of the state-of-the-art solutions that closely align with our contribution. Section 4 presents the proposed method. In Section 5, we report several examples testing different datasets to demonstrate our proposal's novelty and effectiveness for counterfactual generation. Finally, Section 6 indicates future research directions and discusses the limitations of our approach.

## 2. Related Work

### 2.1. Counterfactual generation techniques

Counterfactual approaches explain individual predictions describing what-if contrastive scenarios [4,5,6]. 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.

Various approaches have been devised for generating counterfactuals, including optimization, heuristic search, instance-based, and decision trees. We refer to [7] for a more comprehensive overview. 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* [8,9,10]. The complexity of this range of evaluation metrics presents a critical challenge in comparing the currently available algorithms and establishing benchmark procedures. Moreover, most of the current counterfactual generation techniques do not allow for user interaction, thus limiting the practical applicability of these methods in real-world contexts [11].

### 2.2. Human-Centered Approach

It is crucial to rethink new ways of interaction between humans and the XAI algorithm, focusing more on the user's point of view and needs [12,13,14,15]. The effectiveness of an explanation is based on the extent to which the recipient regards it as comprehensible and useful. A human-centered approach is needed to fill the gap between the algorithmic perspective and an explanation that is truly intelligible to humans [13]. In fact, as stated Liao et al. [13], XAI presents more of a design challenge than an algorithmic challenge.

Previous social science research has shown that human explanations are often contrastive. Therefore, counterfactuals are generally considered less cognitively demanding and more intuitive [16]. Given these assumptions, our proposed interface and algorithm are outlined as starting points for several case studies, see Section 5, and are therefore designed for domain experts and lay users [17].

In 2018 the DARPA explainable AI initiative framed the explainable AI process as a three-stage approach, distinguishing between the explainable model, the explanation user interface. [18]. This definition draws a difference between the model through which an explanation of a machine learning algorithm is generated and the means used to communicate it to the user [19]. From this perspective, the explanation process has to be outlined as a continuous dialogue between a sender and a receiver [14]. Therefore XAI has to consider interactivity as a fundamental part of the process [20,12,19], for example, through the design of novel interfaces that allow model inspection at will [21], and promote an understanding of the computational processes giving users control of the algorithmic actions [22].

### *2.3. Interactive explanations and visualizations*

In the following, we frame `Feature Space Navigator` in relation to existing literature solutions, and critically discuss the novelty and positioning of our contribution.

`DECE` is an advanced counterfactual generation method tailored for both global and local interpretability tasks [23]. With a strong focus on visualization and user interface design, `DECE` automates counterfactual generation using established metrics such as diversity and validity. Notably, it facilitates user-driven constraint imposition through a decision tree framework, offering fine-grained control over the generation process. Moreover, `DECE` extends beyond dataset limitations by exploring counterfactuals independent of available data, while leveraging subset counterfactuals to dissect machine learning model decision boundaries.

`Prospector` is a method designed to generate instance counterfactuals via a visual interface [24]. It utilizes a semi-global score, specifically a partial dependence plot, to estimate the probable prediction when specific features of the original instance are modified. This process involves three distinct phases: patient selection, patient inspection, and the generation of partial dependence plots. Access to both the model and the dataset used for training is required. The tool is specifically designed for data scientists. It allows them to understand how features affect the prediction, thus improving a predictive model.

`DiCE` framework generates and evaluates counterfactuals according to two properties, namely *plausibility* and *diversity* [25]. The authors test `DiCE` on four real-world datasets and compare its counterfactuals with popular local explanation methods and other prior approaches.

`T-LACE` generates reliable counterfactual explanations leveraging on two properties of a tailored transparent latent space, namely *similarity* and *linearity* [26]. An interactive framework for auto-encoders allows a visual exploration of the latent space, shedding insights on the links between input features and model prediction [27].

For a thorough overview, we refer to [28], where authors present techniques and strategy designed to explore and explain model's predictions at *instance-level*, i.e., for a single record, and at *dataset-level*, i.e., for an entire collection.

*Our contribution* We outline some of the distinguishing properties of `Feature Space Navigator`, our method, setting it apart from existing literature. Firstly, it prioritizes the user, who not only operates the tool but also indirectly defines the optimization metric, ensuring subjective considerations are captured. Moreover, it requires minimal technical skills, as no coding knowledge is needed, and users interact through intuitive sliders. It works with any model and does not require access to the dataset, making it easy to explore different models without constraints. Furthermore, it is purely local, focusing solely on the specific instance at hand, ensuring relevance to the user's context. Users can adjust instances feature by feature, enabling thorough exploration of the feature space. Overall, our method is user-friendly, adaptable, and effective for exploring machine learning systems.

## 3. Setting the Stage

*Preliminaries* Given a classifier $b$ that outputs the decision $y = b(x)$ for an instance $x$, a counterfactual explainer $\mathscr{E}$ 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 $\ell_p$ norm [11]. In the vast majority of cases $c$ exhibits the following limitations:

**Limitation 1.** *Stationarity*: $c(x,x')$ is predetermined and not updated through user feedback. Moreover, it lacks the ability for $\mathscr{E}$ to adapt during post-deployment usage, which would allow for fine-tuning between user needs and the tool's effectiveness (see *Research Challenge 9* in [29]).

**Limitation 2.** *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£.

**Limitation 3.** *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 (see *Research Challenge 10* in [29]). An attempt to address this limitation can be found in [30].

Our contribution aims to address these limitations, especially from a user-focused viewpoint.

## 4. Feature Space Navigator

Building upon [31], 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 proposal leverages the autonomy and control of the users that can explore the behavior of the decision model accordingly with their own knowledge base, reducing the need for a dedicated explanation algorithm. From this perspective, users are designated as the "domain experts" of their own real-word scenario.

`Feature Space Navigator` is local, post-hoc and agnostic, i.e., designed to account for any black box[2]. It offers an interactive interface of the areas where the model returns the desired outcome by providing a graphical visualization. The entire process is delegated to the users that are free to explore the feature space of the system and generate autonomously, "by hand", effective counterfactual and prototype instances that they deem feasible, desirable, and suitable to their specific evaluations and needs[3]. Users are therefore empowered with complete autonomy and control over the process without the mediation of an explanation algorithm between the human and the decision model. In

[2] `Feature Space Navigator` is available at https://github.com/Elecapp/Feature_space_navigator.

[3] User interactions through the interface are not stored; hence, no concern occurs regarding the safety of private information that users might input.

```
Algorithm 1: GridEvaluation($ranges$, $bins$, $X$, $x$, $model$)
  Input  : $ranges$ - features ranges, $bins$ - number of bins, $X$ - dataset, $x$ - original
           instance, $model$ - black box
  Output: $\hat{Y}$ - $model$'s predictions
  $d \leftarrow \mathrm{len}(x)$;                                   // storing number of features
  $\hat{Y} \leftarrow [0]_{d \times \mathrm{bins}}$ ;               // creating a matrix of zeros
  for $i = 0$ to $d$ do
      if $categorical(x_i)$ then
          for $v \in unique(X_i)$ do
              $\hat{x} \leftarrow \mathrm{copy}(x)$ ;                          // creating a copy of $x$
              $\hat{x}_i \leftarrow v$ ;    // changing $\hat{x}$'s $i$-th feature with the new value
              $\hat{Y}_{i,b} \leftarrow \mathrm{model}(\hat{x})$ ;                  // predicting score
      else
          $min, max \leftarrow ranges_i$ ;   // saving min and max of current feature
          for $b = 0$ to $bins$ do
              $v \leftarrow min + \frac{b}{\mathrm{bins}} \cdot (max - min)$ ;       // computing the new value
              $\hat{x} \leftarrow \mathrm{copy}(x)$ ;                          // creating a copy of $x$
              $\hat{x}_i \leftarrow v$ ;    // changing $\hat{x}$'s $i$-th feature with the new value
              $\hat{Y}_{i,b} \leftarrow \mathrm{model}(\hat{x})$ ;                  // predicting score
  return $\hat{Y}$
```

the following, we outline the interface, the underlying algorithm, and discuss evaluation aspects of the outputs returned by `Feature Space Navigator`.

*Interface* The interface, depicted in Section 5, visually displays how model predictions react in response to the adjustments introduced by the users, letting them to identify relevant prototypes and counterfactuals. It allows users to independently explore the feature space by selectively modifying the original instance on the basis of her own knowledge and experience, i.e., changing the value of one of the features that characterize the instance $x$. This can be done introducing minimal changes through the adjustment of the slider *feature by feature* within its designated range. Specifically, 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. Through a number of attempts, generating both prototypes and counterfactuals, the user intuitively guesses the feature importance of the classifier decision boundary through gradients change.

*Algorithm* The user can move along each axis in the feature space but can not 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 process. It requires access to the model's probability outputs and is specifically tailored for tabular data with a binary target variable. The number of predictions required scales linearly with the number of features, ensuring the interface update remains essentially real-time.

## 5. Scenarios

*Dataset* In this work, we implemented `Feature Space Navigator` on a dataset widely used in academic research and freely accessed through the UCI Repository [32]: the STATLOG (GERMAN CREDIT DATA) dataset [33], which categorizes individuals as having good or bad credit risk through a binary classification task. It comprises 1000 instances and 20 features, all of which are categorical and integer types. The categorical features are handled in the interface by preserving the original categorical feature's record alongside the value generated by the LabelEncoder.

*Classifiers* We performed an 85-15 split in train and test for STATLOG and a 70-30 split for WINE. We ran our evaluation on Random Forest implemented by the `scikit-learn` library[4]. The best parameters are identified through a Randomized Search over a dictionary of possibilities. Both the classifiers were performed using a fixed random state.

*Understanding the Effects of Interventions* In the following, we provide examples of potential usage of the interface on the adopted datasets, aiming to illustrate the functionality of `Feature Space Navigator`.

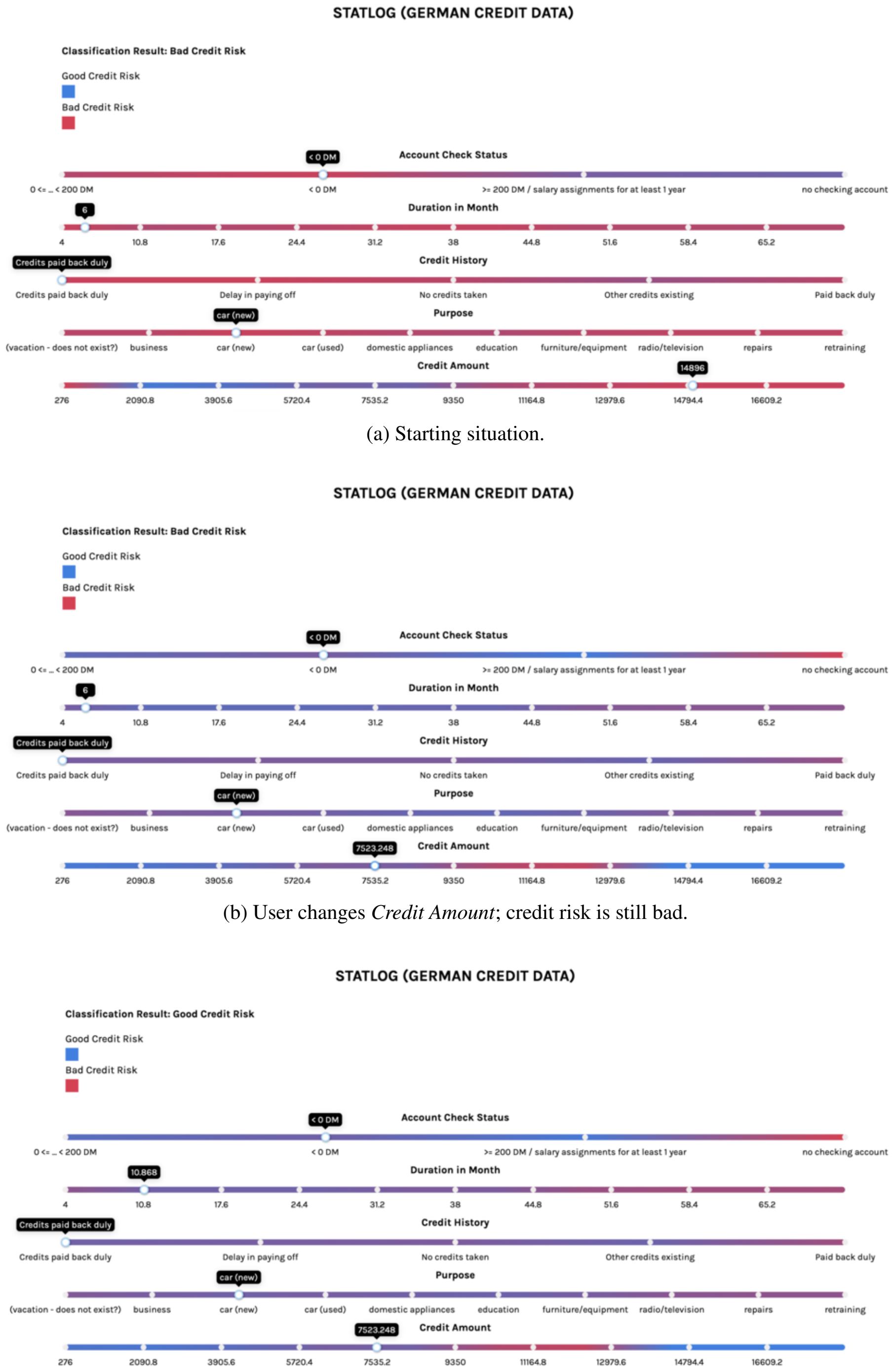
STATLOG *Scenario* Let us first examine a case from the STATLOG dataset, where a user applies for a loan. The initial situation, depicted in Fig. 2a, receives a negative outcome (the credit risk is bad, therefore the loan is not granted). The user, after scrutinizing the interface, decides to lower the *credit amount* requested, as documented in Figure 2b. Since the outcome issued by the black box remains negative, the user is not satisfied and applies a different change, this time adjusting the *duration in month* of the loan (Figure 2c). In this case the user finally receives a good credit risk classification.

## 6. Conclusion

In this work, we have developed `Feature Space Navigator`, an interactive graphical interface that facilitates exploration of the feature space of the decision boundary of a black box, providing an intuitive visualization of the areas where the model produces the desired outcomes. This user-centric approach enhances transparency and control in the decision-making process, contributing to the advancement of XAI methodologies.

Embracing a human-centered perspective, the evaluation requires a conceptual shift by involving users within it in order to understand whether the explanation is convincing for the intended user, the ultimate stakeholder in the process [34,35]. Concretely, a qualitative evaluation in the form of a user study is necessary. The study 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.

[4] `https://scikit-learn.org/stable/index.html`



(a) Starting situation.

(b) User changes *Credit Amount*; credit risk is still bad.

(c) User changes *Duration in Month*; credit risk becomes good.

**Figure 2.** STATLOG Scenario.

## Acknowledgements

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 the NextGenerationEU programme under the funding schemes PNRR-PE-AI scheme (M4C2, investment 1.3, line on AI) *FAIR* (Future Artificial Intelligence Research) Prot. PE00000013.

## References

[1] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM computing surveys (CSUR). 2018;51(5):1-42.

[2] Sovrano F, Vitali F, Palmirani M. Making Things Explainable vs Explaining: Requirements and Challenges under the GDPR. CoRR. 2021;abs/2110.00758. Available from: https://arxiv.org/abs/2110.00758.

[3] Seizov O, Wulf A. Artificial Intelligence and Transparency: A Blueprint for Improving the Regulation of AI Applications in the EU. European Business Law Review. 2020 09;31:611-40.

[4] Karimi A, Schölkopf B, Valera I. Algorithmic Recourse: from Counterfactual Explanations to Interventions. In: FAccT. ACM; 2021. p. 353-62.

[5] Molnar C, Casalicchio G, Bischl B. 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. vol. 1323 of Communications in Computer and Information Science. Springer; 2020. p. 417-31. Available from: https://doi.org/10.1007/978-3-030-65965-3_28.

[6] Beretta I, Cinquini M. The Importance of Time in Causal Algorithmic Recourse. In: Longo L, editor. Explainable Artificial Intelligence - First World Conference, xAI 2023, Lisbon, Portugal, July 26-28, 2023, Proceedings, Part I. vol. 1901 of Communications in Computer and Information Science. Springer; 2023. p. 283-98. Available from: https://doi.org/10.1007/978-3-031-44064-9_16.

[7] Guidotti R. Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery. 2022:1-55.

[8] Jia Y, Bailey J, Ramamohanarao K, Leckie C, Houle ME. Improving the Quality of Explanations with Local Embedding Perturbations. In: Teredesai A, Kumar V, Li Y, Rosales R, Terzi E, Karypis G, editors. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019. ACM; 2019. p. 875-84. Available from: https://doi.org/10.1145/3292500.3330930.

[9] Zhang Y, Song K, Sun Y, Tan S, Udell M. " Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. arXiv preprint arXiv:190412991. 2019.

[10] Guidotti R. Evaluating local explanation methods on ground truth. Artif Intell. 2021;291:103428. Available from: https://doi.org/10.1016/j.artint.2020.103428.

[11] Karimi A, Barthe G, Schölkopf B, Valera I. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. ACM Comput Surv. 2023;55(5):95:1-95:29. Available from: https://doi.org/10.1145/3527848.

[12] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence. 2019;267:1-38.

[13] Liao QV, Varshney KR. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. CoRR. 2021;abs/2110.10790. Available from: https://arxiv.org/abs/2110.10790.

[14] Abdul AM, Vermeulen J, Wang D, Lim BY, Kankanhalli MS. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In: Mandryk RL, Hancock M, Perry M, Cox AL, editors. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018. ACM; 2018. p. 582. Available from: https://doi.org/10.1145/3173574.3174156.

[15] Amershi S, Cakmak M, Knox WB, Kulesza T. Power to the People: The Role of Humans in Interactive Machine Learning. AI Mag. 2014;35(4):105-20. Available from: https://doi.org/10.1609/aimag.v35i4.2513.

[16] Riveiro M, Thill S. ”That’s (not) the output I expected!” On the role of end user expectations in creating explanations of AI systems. Artif Intell. 2021;298:103507. Available from: https://doi.org/10.1016/j.artint.2021.103507.

[17] Ribera M, Lapedriza À. Can we do better explanations? A proposal of user-centered explainable AI. In: Trattner C, Parra D, Riche N, editors. Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI 2019), Los Angeles, USA, March 20, 2019. vol. 2327 of CEUR Workshop Proceedings. CEUR-WS.org; 2019. Available from: https://ceur-ws.org/Vol-2327/IUI19WS-ExSS2019-12.pdf.

[18] Gunning D, Aha D. DARPA’s explainable artificial intelligence (XAI) program. AI magazine. 2019;40(2):44-58.

[19] Chromik M, Schuessler M. A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. Exss-atec@ iui. 2020;94.

[20] Madumal P, Miller T, Sonenberg L, Vetere F. A grounded interaction protocol for explainable artificial intelligence. arXiv preprint arXiv:190302409. 2019.

[21] Shneiderman B. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. ACM Trans Interact Intell Syst. 2020;10(4):26:1-26:31. Available from: https://doi.org/10.1145/3419764.

[22] Shneiderman B, Plaisant C, Cohen MS, Jacobs S, Elmqvist N, Diakopoulos N. Grand challenges for HCI researchers. Interactions. 2016;23(5):24-5. Available from: https://doi.org/10.1145/2977645.

[23] Cheng F, Ming Y, Qu H. DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. IEEE Trans Vis Comput Graph. 2021;27(2):1438-47. Available from: https://doi.org/10.1109/TVCG.2020.3030342.

[24] Krause J, Perer A, Ng K. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In: Kaye J, Druin A, Lampe C, Morris D, Hourcade JP, editors. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016. ACM; 2016. p. 5686-97. Available from: https://doi.org/10.1145/2858036.2858529.

[25] Mothilal RK, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. In: Hildebrandt M, Castillo C, Celis LE, Ruggieri S, Taylor L, Zanfir-Fortuna G, editors. FAT* ’20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020. ACM; 2020. p. 607-17. Available from: https://doi.org/10.1145/3351095.3372850.

[26] Bodria F, Guidotti R, Giannotti F, Pedreschi D. Transparent Latent Space Counterfactual Explanations for Tabular Data. In: Huang JZ, Pan Y, Hammer B, Khan MK, Xie X, Cui L, et al., editors. 9th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2022, Shenzhen, China, October 13-16, 2022. IEEE; 2022. p. 1-10. Available from: https://doi.org/10.1109/DSAA54385.2022.10032407.

[27] Bodria F, Rinzivillo S, Fadda D, Guidotti R, Giannotti F, Pedreschi D. Explaining Black Box with Visual Exploration of Latent Space. In: Agus M, Aigner W, Höllt T, editors. 24th Eurographics Conference on Visualization, EuroVis 2022 - Short Papers, Rome, Italy, June 13-17, 2022. Eurographics Association; 2022. p. 85-9. Available from: https://doi.org/10.2312/evs.20221098.

[28] Biecek P, Burzykowski T. Explanatory Model Analysis. Chapman and Hall/CRC, New York; 2021. Available from: https://pbiecek.github.io/ema/.

[29] Verma S, Dickerson JP, Hines K. Counterfactual Explanations for Machine Learning: A Review. CoRR. 2020;abs/2010.10596. Available from: https://arxiv.org/abs/2010.10596.

[30] Mahajan D, Tan C, Sharma A. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. CoRR. 2019;abs/1912.03277. Available from: http://arxiv.org/abs/1912.03277.

[31] Beretta I, Cappuccio E, Manerba MM. User-Driven Counterfactual Generator: A Human Centered Exploration. In: Longo L, editor. Joint Proceedings of the xAI-2023 Late-breaking Work, Demos and Doctoral Consortium co-located with the 1st World Conference on eXplainable Artificial Intelligence (xAI-2023), Lisbon, Portugal, July 26-28, 2023. vol. 3554 of CEUR Workshop Proceedings. CEUR-WS.org; 2023. p. 83-8. Available from: https://ceur-ws.org/Vol-3554/paper15.pdf.

[32] Kelly M, Longjohn R, Nottingham K. The UCI Machine Learning Repository;. https://archive.ics.uci.edu.

[33] Hofmann H. Statlog (German Credit Data); 1994. DOI: https://doi.org/10.24432/C5NC77. UCI Machine Learning Repository.

[34] Liao QV, Zhang Y, Luss R, Doshi-Velez F, Dhurandhar A. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. In: Hsu J, Yin M, editors. Proceedings of the Tenth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2022, virtual, November 6-10, 2022. AAAI Press; 2022. p. 147-59. Available from: https://ojs.aaai.org/index.php/HCOMP/article/view/21995.

[35] Wang ZJ, Vaughan JW, Caruana R, Chau DH. GAM Coach: Towards Interactive and User-centered Algorithmic Recourse. In: Schmidt A, Väänänen K, Goyal T, Kristensson PO, Peters A, Mueller S, et al., editors. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023. ACM; 2023. p. 835:1-835:20. Available from: https://doi.org/10.1145/3544548.3580816.