ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Reliable Counterfactuals for Machine Learning Models – Current Aspects and Perspectives

Marika Kaden[1], Benjamin Paaßen[2], Ronny Schubert[1], Barbara Hammer[2], and Thomas Villmann[1,3] *

1 - University of Applied Sciences Mittweida, Saxon Institute for Comp. Intelligence and Machine Learning, Mittweida - Germany

3 - University Bielefeld, Center for Cognitive Interaction Technology, Bielefeld - Germany

3 - Technical University Bergakademie Freiberg, Freiberg - Germany

Abstract. Counterfactuals are becoming increasingly important for explaining and evaluating machine learning approaches. In crucial and challenging applications such as medical diagnosis support, credit scoring, and technical control systems, evaluating counterfactuals is a promising methodology to exploring the applicability and limits of AI systems. This approach also empowers users with actionable advice on how to influence automatic decisions. Thus, reliable and faithful generation as well as a prudent interpretation of counterfactuals contribute to establishing more reliable AI systems and improving their trustworthiness.

Determination and generation of counterfactual samples can be motivated and processed taking two different perspectives: The cognitive perspective considers trustworthiness and reliance based on empirical evidence in human reasoning or model explanations inspired by experiences in social sciences. The technical perspective is mainly triggered by issues like plausibility and actionability of counterfactuals as well as by their efficient computation and evaluation. Current developments in counterfactual research provide substantial progress but are far away from to be sufficient in the field.

In this introduction paper, we highlight some current aspects in this interdisciplinary field of research inspired by cognitive models of inference and reasoning as well as triggered by technical developments in the field of machine learning and artificial intelligence.

1 Introduction

Counterfactuals constitute an important paradigm for human inference mechanisms and learning intensively studied in cognitive and social sciences [42, 50, 47], as well as in cognitive learning theory [21]. Simply speaking, counterfactuals and counterfactual thinking provide thoughts about alternatives to past events,

*M.K. and R.S. are financially supported by the network project KI-Med founded by the European Social Fund (ESF, 100734114). B.P. and B.H. are supported by the Federal Ministry of Research, Technology and Space (BMFTR) under grant no. 01IS24057A (KI-Akademie OWL).

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

meaning thoughts of what might have been [25]. The traditional paradigm for counterfactual reasoning in this literature is the interventional counterfactual, where hypothetical interventions are imagined and simulated [54, 56, 66]. In this sense, counterfactual reasoning makes an obligatory contribution also to learning, by providing alternative scenarios from which to draw conclusions.

In machine learning, counterfactuals play a key role for model explanation in the context of explainable and interpretable AI systems [18, 29, 48, 61, 51]. To achieve this explanation behavior, counterfactual generation methods can benefit from the social and cognitive sciences [15, 16, 45]. Further, counterfactuals can be used to design causal machine learning approaches [29, 46], or to evaluate causal inference schemes [11, 20].

Several strategies and methods are known to generate counterfactuals [4, 18, 52, 61]. However, we have to carefully distinguish between counterfactual generation and adversarial attacks,as they are similar yet differ in important aspects. Counterfactuals are used to explain a model, which will be discussed in more detail later. Adversarial examples, on the other hand, try to fool the model.[49, 24].

In this paper we briefly highlight current developments and challenges regarding generation or determination of counterfactuals and their use for model explanations and derived inference schemes. We situate contributions to the ESANN-special-session 'Reliable Counterfactuals for Machine Learning Models' in this context.

The paper is organized as follows: After this initial chapter, we start with a section introducing the basic concepts regarding counterfactual investigations, generations and explanations taking the perspective of machine learning. Thereafter, we point out several aspects of counterfactual generation and explanation as well as evaluation followed by a perspective view for future research challenges and directions. Brief concluding remarks summarize future directions of research and respective challenges.

2 Counterfactuals in Machine Learning – Basic Concepts

In machine learning, a counterfactual explanation describes how a data point would need to change to receive a different classification. More formally, we suppose a classifier model $M : \mathbb{R}^n \supseteq \mathcal{D} \ni \mathbf{x} \mapsto c_M(\mathbf{x}) \in \mathcal{C}$ where $c_M(\mathbf{x})$ is a class label predicted by the model M and contained in the class set $\mathcal{C}$, and $\mathcal{D}$ is the (usually unknown) data manifold. In case of training data, also the true class label $c(\mathbf{x}) \in \mathcal{C}$ is available. Further, a dissimilarity measure d is required to compare data. Following the standard formulation for counterfactuals proposed in [67], a counterfactual sample $\mathbf{x}_{\text{cf}} \in \mathbb{R}^n$ for given $\mathbf{x}$ is determined by

$$\delta_{\text{cf}}(\mathbf{x}, \mathbf{x}_{\text{cf}}) \stackrel{!}{=} \min_{\tilde{\mathbf{x}} \in \mathcal{D}} (d(\mathbf{x}, \tilde{\mathbf{x}})) \quad \text{subject to} \quad c_M(\mathbf{x}) \neq c_M(\mathbf{x}_{\text{cf}}), \tag{1}$$

constituting a constrained optimization problem. It should be emphasized that the constraint is given by the comparison of the model predictions $c_M(\mathbf{x})$ and

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

original instance $\mathbf{x}$	adversarial sample $\mathbf{x}_{\text{ae}}$	counterfactual sample $\mathbf{x}_{\text{cf}}$
objective	$c(\mathbf{x}) \neq c_M(\mathbf{x}_{\text{ae}})$	$c_M(\mathbf{x}) \neq c_M(\mathbf{x}_{\text{cf}})$
sample variation	$\delta_{\text{ae}}(\mathbf{x}, \mathbf{x}_{\text{ae}}) < \epsilon$	$\delta_{\text{cf}}(\mathbf{x}, \mathbf{x}_{\text{cf}}) \overset{!}{=} \min$
goal	fool the model	explore/explain the model limits
strategy	may use adversarial dissimilarities δ_{ae} difficult to distinguish by human perception	should use dissimilarities δ_{cf} being consistent with the model

Table 1: Comparison adversarial sample versus counterfactual sample

$c_M(\mathbf{x}_{\text{cf}})$ for the given sample $\mathbf{x}$ and its corresponding counterfactual $\mathbf{x}_{\text{cf}}$. If counterfactuals should be used for model explanations like in explainable artificial intelligence (XAI, [57]), the dissimilarity measure d has to be chosen consistently to the model. For example, (deep) multilayer perceptrons implicitly make use of the Euclidean distance due to the perceptron units, which are based on the Euclidean inner product. Otherwise, kernel methods use kernel distances taking into account the inner-product-property of kernels [59, 65]. Further, if a certain class prediction $c^*(\mathbf{x}_{\text{cf}}) \neq c_M(\mathbf{x})$ is desired for the counterfactual determination, this task is denoted as *guided counterfactual generation* (GCG). The GCG is of particular interest, if multiple model explanations regarding different aspects are demanded.

The counterfactual sample $\mathbf{x}_{\text{cf}}$ has to be carefully distinguished from adversarial examples $\mathbf{x}_{\text{ae}}$, which are designed to fool the model in such a way that the model class assignment $c_M(\mathbf{x}_{\text{ae}})$ differs from the *true* sample class $c(\mathbf{x})$ with a small deviation $\delta_{\text{ae}}(\mathbf{x}, \mathbf{x}_{\text{ae}}) < \epsilon$ for given threshold $\epsilon > 0$ [24, 49]. In particular, the adversarial deviation measure δ_{ae} is usually selected in such a way that in this adversarial aim small disturbances of the original $\mathbf{x}$ contained in the adversarial $\mathbf{x}_{\text{ae}}$ are difficult to detect by human perception (see also Table 1). However, the internal (perhaps implicit) model dissimilarity usually becomes large, leading to model misclassification $c_M(\mathbf{x}_{\text{ae}}) \neq c(\mathbf{x})$ [26, 24, 49].

Paradigms and methods for counterfactual generation as well as respective evaluation criteria are considered in the next section.

3 Reliable Counterfactuals

The generation of counterfactuals follows different perspectives and strategies including efficient generalization approaches and evaluation properties. In particular, the following aspects can be identified to achieve reliable counterfactuals:

- **Validity**: the counterfactual $\mathbf{x}_{\text{cf}}$ is taken as a perturbation of the original sample $\mathbf{x}$, i.e. $\mathbf{x}_{\text{cf}} = \mathbf{x} + \Delta_G(\mathbf{x})$ preferably realized by a mathematically well-defined perturbation generator function $G : \mathbf{x} \mapsto \Delta_G(\mathbf{x})$

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

- **Proximity**: keeps the counterfactual $\mathbf{x}_{\text{cf}}$ close to the original sample $\mathbf{x}$
- **Sparsity**: aims to minimize the complexity of the perturbation, e.g. minimization of $\|\Delta_G(\mathbf{x})\|_0$
- **Plausibility**: requires the counterfactuals to be realistic, i.e. they should belong to the (usually unknown) data manifold $\mathcal{D}$
- **Actionability/Feasibility**: the perturbation $\Delta_G(\mathbf{x})$ must also be truly feasible

Beside these qualitative features, efficient determination of counterfactuals is an important aspect, particularly if many different aspects have to be taken additionally into account regarding specifically given circumstances. Due to the given optimization problem (1), the majority of counterfactual generating approaches deal with efficient realizations of the optimization task [37]. For example, in case of convex problem formulation of (1), the optimization procedure becomes fast [4, 37] whereas zero-shot generation guided by Large Language Models (LLM) are proposed in [13] the context of natural language processing. As pointed out in [33], a geometric-analytical determination is possible for the special GCG using vector quantization classifiers [14], if the dissimilarity measure d is induced by an inner product, for example the squared Euclidean distance. Model agnostic approaches are studied in [7, 43, 34].

Closely related to the proximity quality of counterfactuals is the challenge of their (numerical) robustness and of the respective explanations [30, 31, 32, 40, 41], as well as its evaluation [6]. Yet, this problem often cannot be considered without investigation of the underlying machine learning model [17]. In case of neural networks, probabilistic guarantees and efficient computation strategies can be realized [27, 4]. In [64] counterfactual probabilities are determined using bivariate distributions and uplift modeling. More generally, robustness and diversity is addressed in [40, 41]. Another important aspect for this perspective is that those approaches can benefit from data feature importance as provided by Shapley-approaches [1, 55]

Plausible counterfactuals are mainly understood as those counterfactual candidates which should belong to the data manifold $\mathcal{D}$ and, hence, leads to improved trustworthiness [18, 36, 44]. In other words: a counterfactual example generated by the state-of-the-art systems is not necessarily representative of the underlying data distribution $\mathcal{D}$ and may therefore prescribe *unachievable goals* which has to be avoided for practical applications. Accordingly, approaches for plausible counterfactuals determination usually are based on the estimation of the empirical class-dependent data density required to cover counterfactual candidate regions [3, 4].

Further, the perturbation $\Delta_G(\mathbf{x})$ does not always realize a "feasible path" between the current state $\mathbf{x}$ and the generated counterfactual $\mathbf{x}_{\text{cf}}$, making actionable recourse infeasible. In [52] an algorithmically sound way of uncovering these "feasible paths" is proposed, which is based on the shortest path distances

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

defined via density-weighted metrics. This approach is claimed to generate counterfactuals being coherent with the underlying data manifold $\mathcal{D}$ and, therefore, are achievable and can be tailored to the problem at hand. Additionally we emphasize that the feasibility property has to be consistent with the deviation measure $\delta_{\text{cf}}(\mathbf{x}, \mathbf{x}_{\text{cf}}) = \delta_{\text{cf}}(\mathbf{x}, \mathbf{x} + \Delta_G(\mathbf{x}))$ from (1) for counterfactual generation. To be more precisely, the corresponding dissimilarity measure d for the data manifold $\mathcal{D}$ has chosen carefully to reflect the feasibility aspect.

Yet, plausibility can also be considered from a user-centric view [38]. Further, taking a cognitive science perspective one can think how human subjects perceive counterfactual explanation [15, 63].

Obviously, plausibility of counterfactuals also relates to validity and classification robustness of the model. In this context, reject options (or *abstention*) and methods for classification models are of interest [23, 39, 53]. The rejection policy is typically formalized as a function where if the distance to the nearest counterfactual is below a threshold, the input is rejected [62]. For them, special counterfactuals known as *semi-factual explanations* ('even if' – statements, [5, 8, 22]) are proposed. In contrast to usual counterfactuals suggesting actions for change ('if only' – statement), semi-factuals helping users to understand that a certain result was 'bound to happen' regardless of small input changes and, hence, can be seen as explanations why a model refuses to make a decision or declines a suggested prediction. Further, semi-factual methods can be categorized as counterfactual-guided (using the closest, opposite-class instance) or counterfactual-free [9]. Current investigations deal with the question, whether best semi-factual explanations can be found using counterfactuals as guides [10].

Sparsity oriented counterfactual generation keeping the diversity is studied in [28] using gradient-based methods. This strategy usually takes a latent representation space into account to generate multiple counterfactual explanations that are sparse, realistic, and robust to input manipulations [58]. Otherwise, sparseness for counterfactuals can be enforced by respective feature sparseness of $\Delta_G(\mathbf{x})$ using the game-theoretic feature attributions known as Shapley-values [1, 2].

4 Concluding Remarks

Overall, the determination and evaluation of counterfactuals remains challenging and depends from the application case and usually has to incorporate domain knowledge [60]. Thus different areas require different strategies and methods for efficient counterfactual generation, explanation and evaluation. For example, in [12] graph diffusion counterfactual explanations are proposed, which may offer new application areas of counterfactuals for relational or structured data as known to be occurring in many research areas as molecular bioinformatic, social-sciences and others.

Related to this application perspective, also the challenge of feasibility in relation to the appropriate choice of the manifold dissimilarity becomes more and more important to achieve trustworthiness for counterfactual explanations.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Therefore, explicit domain knowledge integration might offer a possibility to deal with this problem.

Further, contrastive approaches as used in algorithmic recourse and consequential recommendations as well as other cognitive learning and reasoning strategies could provide further inspiration for future research in the field of counterfactuals and their explanation [19, 35, 55, 61].

References

[1] E. Albini, J. long, D. Dervovic, and D. Magazzeni. Counterfactual Shapley additive explanations. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT 2022)*, pages 1054–1070, 2022.

[2] E. Albini, S. Sharma, S. Mishra, D. Dervovic, and D. Magazzeni. On the connection between game-theoretic feature attributions and counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2023:)*, pages 411–433. Association for Computing Machinery (ACM), 2023.

[3] A. Artelt and B. Hammer. Convex density constraints for computing plausible counterfactual explanations. In I. Farkaš, P. Masulli, and S. Wermter, editors, *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, volume 12396 of *LNCS*, pages 353–363, 2020.

[4] A. Artelt and B. Hammer. Efficient computation of counterfactual explanations and counterfactual metrics of prototype-based classifiers. *Neurocomputing*, 470:304–317, 2022.

[5] A. Artelt and B. Hammer. "even if ..." – diverse semifactual explanations of reject. In *Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 854–859. IEEE Press, 2022.

[6] A. Artelt, V. Vaquet, R. Velioglu, F. Hinder, J. Brinkrolf, M. Schilling, and B. Hammer. Evaluating robustness of counterfactual explanations. In *Proceedings of IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–9. IEEE Press, 2021.

[7] A. Artelt, R. Visser, and B. Hammer. "i do not know! but why?" – local model-agnostic example-based explanations of reject. *Neurocomputing*, 558(126722):1–11, 2023.

[8] S. Aryal. Semi-factual explanations in AI. In *Proceedings of the Doctoral Consortium at ICCBR 2024*, volume 3708 of *CEUR-WS Proceedings*, pages 236–240. CEUR-WS.org, 2024.

[9] S. Aryal and M. Keane. Even if explanations: Prior work, desiderata and benchmarks for semi-factual XAI. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023)*, pages 6526–6535, 2023.

[10] S. Aryal and M. Keane. Even-Ifs from If-Onlys: Are the best semi-factual explanations found using counterfactuals as guides? In J. Recio-Garcia, M. O. del Castillo, and D. Bridge, editors, *Proceedings of the 32nd International Conference on Case-Based Reasoning Research and Development (ICCBR 2024)*, number 14775 in Lecture Notes in Artificial Intelligence (LNAI), pages 33–49. Springer Nature, 2024.

[11] S. Baron. Explainable AI and causal understanding: Counterfactual approaches considered. *Minds and Machines*, 33:347–377, 2023.

[12] D. Bechtholdt and S. Bender. Graph diffusion counterfactual explanation. In M. Verleysen, editor, *Proceedings of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2026), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2026. i6doc.com.

[13] A. Bhattacharjee, R. Moraffahand, J. Garland, and H. Liu. Zero-shot LLM-guided counterfactual generation: A case study on NLP model evaluation. In *Proceedings of the IEEE International Conference on Big Data (BigData)*, pages 1243–1248, 2024.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

[14] M. Biehl, B. Hammer, and T. Villmann. Prototype-based models in machine learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 7(2):92–111, 2016.

[15] R. Byrne. Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19, Macao)*, pages 6276–6282, 2019.

[16] L. Celar and R. Byrne. How people reason with counterfactual and causal explanations for Artificial Intelligence decisions in familiar and unfamiliar domains. *Memory and Cognition*, 51:1481–1496, 2023.

[17] L. Christodoulou and C. Sun. The impact of machine learning uncertainty on the robustness of counterfactual explanations. *Expert Systems with Applications*, 309(p131198):1–17, 2026.

[18] J. Del Ser, A. Barredo-Arrieta, N. Díaz-Rodríguez, F. Herrera, A. Saranti, and A. Holzinger. On generating trustworthy counterfactual explanations. *Information Sciences*, 655:119898, 2024.

[19] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, and K. Shanmugan. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In S. Bengio, H. W. H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS)*, volume 31 of *Advances in neural information processing systems*, pages 590–601, 2018.

[20] T. Duong, Q. Li, and G. Xu. Causality-based counterfactual explanation for classification models. *Knowledge-Based Systems*, 300(112200), 2024.

[21] K. Epstude and N. Roese. The functional theory of counterfactual thinking. *Personality and Social Psychology Review*, 12(2):168–192, 2008.

[22] O. Espino, I. Orenes, and S. Moreno-Rios. Inferences from the negation of counterfactual and semifactual conditionals. *Memory and Cognition*, 30(5):1090–1102, 2022.

[23] L. Fischer, B. Hammer, and H. Wersing. Efficient rejection strategies for prototype-based classification. *Neurocomputing*, 169:334–342, 2015.

[24] T. Freiesleben. The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines*, 32:77–109, 2021.

[25] D. Galles and J. Pearl. An axiomatic characterization of causal counterfactuals. *Foundation of Science*, 3:151–182, 1998.

[26] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680, San Diego, 2014. Curran Associates, Inc.

[27] F. Hamman, E. Noorani, S. Mishra, D. Magazzeni, and S. Dutta. Robust counterfactual explanations for neural networks with probabilistic guarantees. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning research (PMLR)*, pages 12351–12367, 2023.

[28] C. Han and K. Lee. Gradient-based counterfactual generation for sparse and diverse counterfactual explanations. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing (SAC 2023), Tallin, Estonia*, pages 1240–1247. Association for Computing Machinery (ACM), 2023.

[29] M. Höfler. Causal inference based on counterfactuals. *BMC Medical Research Methodology*, 5(28):1–12, 2005.

[30] J. Jiang, J. Lan, F. Leofante, A. Rago, and F. Toni. Provably robust and plausible counterfactual explanations for neural networks via robust optimisation. In *Proceedings of the 15th Asian Conference on Machine Learning (ACML 2023)*, volume 222 of *Proceedings of Machine Learning Research (PMLR)*, pages 582–597, 2023.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

[31] J. Jiang, F. Leofante, A. Rago, and F. Toni. Formalising the robustness of counterfactual explanations for neural networks. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, number 1671, pages 14901–14909. Association for the Advancement of Artificial Intelligence (AAAI), AAAI press, 2023.

[32] J. Jiang, F. Leofante, A. Rago, and F. Toni. Robust counterfactual explanations in machine learning: A survay. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI-24)*, pages 8086–8094, 2024.

[33] M. Kaden, L. Reuss, and T. Villmann. Geometric-analytical generation of counterfactuals for prototype-based classifiers. In M. Verleysen, editor, *Proceedings of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2026), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2026. i6doc.com.

[34] A.-H. Karimi, G. Bathe, B. Balle, and I. Valera. Model-agnostic counterfactual explanations for consequential decisions. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research (PMLR)*, pages 895–905, 2020.

[35] A.-H. Karimi, G. Bathe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: Contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

[36] E. Kenny and M. Keane. On generating plausible counterfactual and semi-factual explanations for deep learning. In *Proc. Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, pages 11575–11585, 2021.

[37] D. Kirilenko, P. B. M. Gjoreski, M. Luštrek, and M. Langheinrich. Generative models for counterfactual explanations. In *Proccedings of the Human-Interpretable AI Workshop at the KDD 2024 (HI-AI 2024)*, volume 3841 of *CEUR Workshop Proceedings*, pages 18–30. CEUR-WS.org, 2024.

[38] U. Kuhl, A. Artelt, and B. Hammer. Keep your friends close and your counterfactuals closer: Improved learning from ccosest rather than plausible counterfactual explanations in an abstract setting. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 2125–2137, 2022.

[39] T. Landgrebe, D. Tax, P. Paclìk, and R. Duin. The interaction between classification and reject performance for distance-based reject-option classifiers. *Pattern Recognition Letters*, 27:908–917, 2006.

[40] F. Leofante and N. Potyka. Promoting counterfactual robustness through diversity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21322–21329, 2024.

[41] F. Leofante and M. Wicker. Robustness of counterfactual explanations. In *Robust Explainable AI*, SpringerBriefs in Intelligent Systems (BRIEFSINSY), pages 17–40, 2025.

[42] D. Lewis. *Counterfactuals*. Blackwell Publishers, Oxford, 1973.

[43] J. Liu, X. Wu, S. Liu, and S. Gong. Model-agnostic counterfactual explanation: A feature weights-based comprehensive causal multi-objective counterfactual framework. *Expert Systems with Applications*, 266(p126063):1–18, 2025.

[44] F. D. B. Matthew L Stanley, Gregory W Stewart. Counterfactual plausibility and comparative similarity. *Cognitive Science*, 41(5):1216–1228, 2017.

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

[46] S. Morgan and C. Winship. *Counterfactuals and Causal Inference*. Cambridge University Press, 2nd edition, 2014.

[47] S. Morgan and C. Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, 2nd edition, 2014.

[48] R. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proc. Conference on Fairness, Accountability, and Transparency (FAccT), Barcelona*, pages 607–617. Association for Computing Machinery (ACM), 2020.

[49] M. Pawelczyk, C. Agarwal, S. Joshi, S. Upadhyay, and H. Lakkaraju. Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In G. Camps-Valls, F. Ruiz, and I. Valera, editors, *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 151 of *Proceedings of Machine Learning Research*, pages 4574–4594, 2022.

[50] J. Pearl. Causal inference. In *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, volume 6 of *Journal of Machine Learning Research*, pages 39–58, 2010.

[51] J. Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications ACM*, 62(3):54–60, 2019.

[52] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. Bie, and P. Flach. FACE: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pages 344–350. Association for Computing Machinery (ACM), 2020.

[53] C. Punzi, R. Pellungrini, and F. Giannotti. L2loRe: A method for explaining the reject option. In *Proceedings of the Discovery Science Late Breaking Contributions (DS-LB 2024) co-located with 27th International Conference Discovery Science (DS 2024)*, volume 3928 of *CEUR Workshop Proceedings*, pages 1–5. CEIR-WS.org, 2024.

[54] D. Rajamanickam. *Causal Inference for Machine Learning Engineers*, chapter Interventions and Counterfactuals, pages 69–79. Springer Nature, 2026.

[55] S. Rathi. Generating counterfactual and contrastive explanations using SHAP, 2019.

[56] L. Rips and B. Edwards. Inference and explanation in counterfactual reasoning. *Cognitive Science*, 37(6):1107–1135, 2013.

[57] W. Samek, G. Montavon, S. Lapuschkin, C. Anders, and K.-R. Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

[58] S. Sharma, A. Gee, J. Henderson, and J. Ghosh. Faster-ce: Fast, sparse, transparent, and robust counterfactual explanations. In I. Maglogiannis, L. Iliadis, J. Macintyre, M. Avlonitis, and A. Papaleonidas, editors, *Proceedings of Conference Artificial Intelligence Applications and Innovations (AIAI)*, volume 714 of *IFIP Advances in Information and Communication Technology*, pages 183–196, Cham, 2024. Springer Nature Switzerland.

[59] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis and Discovery*. Cambridge University Press, Cambridge, 2004.

[60] N. Spreitzer, H. Haned, and I. van der Linden. Evaluating the practicality of counterfactual explanations. In *Proceedings of the Italian Workshop on Explainable Artificial Intelligence (XAI.it 2022)*, volume 3277 of *CEUR Workshop Proceedings*, pages 31–50. CEUR-WS.org, 2022.

[61] I. Stepin, J. Alonso, A. Catala, and M. Pereira-Farina. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

[62] A. Vailaya and A. Jain. Reject option for VQ-based Bayesian classification. In *International Conference on Pattern Recognition (ICPR)*, pages 2048–2051, 2000.

[63] N. van Hoeck, P. Watson, and A. Barbey. Cognitive neuroscience of human counterfactual reasoning. *Frontiers in Human Neuroscience*, 9(420):1–18, 2015.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

[64] T. Verhelst and G. Bontempi. Identifying counterfactual probabilities using bivariate distributions and uplift modeling. In M. Verleysen, editor, *Proceedings of the 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2026), Bruges (Belgium)*, page in this volume, Louvain-La-Neuve, Belgium, 2026. i6doc.com.

[65] T. Villmann, S. Haase, and M. Kaden. Kernelized vector quantization in gradient-descent learning. *Neurocomputing*, 147:83–95, 2015.

[66] J. von Kügelen, A. Mohamed, and S. Beckers. Backtracking counterfactuals. In *Proceedings of the 2nd Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings Machine learning Research (PMLR)*, pages 1–20, 2023.

[67] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Havard Journal of Law and Technology*, 31(2):841–887, 2018.