# Counterfactual Explanations of Machine Learning Predictions: Opportunities and Challenges for AI Safety

**Kacper Sokol** and **Peter Flach**
Intelligent Systems Laboratory
University of Bristol, UK
{K.Sokol, Peter.Flach}@bristol.ac.uk



## Abstract

One necessary condition for creating a safe AI system is making it transparent to uncover any unintended or harmful behaviour. Transparency can be achieved by explaining predictions of an AI system with counterfactual statements, which are becoming a *de facto* standard in explaining algorithmic decisions. The popularity of counterfactuals is mainly attributed to their compliance with the "right to explanation" introduced by the European Union's General Data Protection Regulation and them being understandable by a lay audience as well as domain experts. In this paper we describe our experience and the lessons learnt from explaining decision tree models trained on UCI German Credit and FICO Explainable Machine Learning Challenge data sets with class-contrastive counterfactual statements. We review how counterfactual explanations can affect an artificial intelligence system and its safety by investigating their risks and benefits. We show example explanations, discuss their strengths and weaknesses, show how they can be used to debug the underlying model, inspect its fairness and unveil security and privacy challenges that they pose.

## Introduction

Safety of software-based artificial intelligence (AI) systems can be achieved in multiple ways. The most fundamental one is to gain users' trust by operating it in a transparent and interpretable way. In turn, this can be achieved by requiring the AI system to explain its actions and decisions to the user. Notwithstanding advantages such as fostering understanding of how a system works, being a tool to debug an AI system and providing a mechanism to inspect its fairness, explanations can also cause unintentional security and privacy vulnerabilities that may compromise the safety of an AI agent. In this paper we show the advantages of counterfactual explanations that can help to improve the overall safety of an AI system and examine security and privacy threats associated with them in a financial domain.

Whether due to the financial market regulations or algorithmic transparency and safety expected by the society, the need for fairness, accountability and transparency of automated decisions in financial services is undisputed. In this work, we examine safety aspects of explaining AI-based loan application and credit scoring systems that use decision trees with class-contrastive counterfactual statements that, usually, are expressed in the following form:

> "The prediction is <prediction>. Had a small subset of features been different <foil>, the prediction would have been <counterfactual prediction> instead."

We have used such counterfactual explanations with predictive AI systems trained on two data sets: UCI German Credit[1] – assessing credit risks based on applicant's personal details and lending history, and FICO Explainable Machine Learning (ML) Challenge[2] – predicting whether an individual has been 90 days past due or worse at least once over a period of 24 months from opening a credit account based on anonymised credit bureau data. To this end, we train decision tree models with scikit-learn (Buitinck et al. 2013) – a Python machine learning library – and generate counterfactual explanations of selected data points with a custom algorithm.

Generating and inspecting some of these statements has provided us with important insights about the limitations, vulnerabilities and properties of counterfactual explanations and their effect on the safety of the underlying AI system. We found that it is important to consider the amount of information that they reveal about the underlying AI system and consequences of this leakage. Moreover, there is a fine line between counterfactual explanations and adversarial examples, which is an important aspect of this approach, especially in a financial setting. One example of a misuse of a counterfactual explanation could be an attempt to game the FICO credit scoring algorithm[3], which is kept as a trade secret to avoid just that. Also, there is usually more than just one counterfactual explanation of the same quality and length, and choosing the most suitable one remains an open research question.

Achieving transparency (explainability and interpretability), fairness and accountability (security and privacy) of AI algorithms, their training data sets and decisions that they output are, in general, open problems. Given the pressure from regulators and society, research in this area has become a hot topic in recent years. New approaches are being

[1] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

[2] https://community.fico.com/s/explainable-machine-learning-challenge

[3] https://www.myfico.com/credit-education/credit-scores/

introduced by researchers and practitioners regularly, either as theoretical considerations or novel algorithms (both open source and commercial). Nevertheless, given historical biases in data (Buolamwini and Gebru 2018) and the fact that most of the predictive algorithms are designed to optimise for predictive performance, their social impact is not always considered. A prominent example of an unexpected vulnerability affecting high-performing and often "well-validated" AI systems are adversarial examples (Nguyen, Yosinski, and Clune 2015). Given that adversarial data points can affect any AI system the danger of gaming predictive algorithms in financial sector is even more concerning.

Counterfactual explanations can help us to address some of these issues. Their most prominent feature is explaining the reasons behind a particular classification outcome (Miller 2019). They can also pick up unfair system behaviour (disparate impact) and unjustified mistreatment of an individual (disparate treatment) (Zafar et al. 2017). Given their concise and easily understandable format they are a useful tool to identify bugs and errors in the underlying predictive model (Kulesza et al. 2015).

Given these advantages, counterfactual explanations are becoming a *de facto* standard in explainable artificial intelligence and their versatility may encourage some product managers to deploy them into intra-company or client-facing applications. Therefore, we decided to investigate advantages, shortcomings, safety issues and potential dangers that can affect a predictive model and its explanations when using counterfactual statements as an explanatory medium. Our findings are presented from a financial data perspective, but should generalise to other safety-critical domains. We hope that these results will prompt the research community and practitioners to consider these issues before deploying explainability techniques in their systems. This is particularly important in the case of sensitive data – e.g. loan applications – or predictive algorithms that should remain secret.

Our observations come from discoveries that we have made while developing and presenting a system that explains decisions of the underlying predictive model with counterfactuals. We have demonstrated our approach to: a lay audience at a local research festival, postgraduate research students at our university and artificial intelligence community attending the 27$^{th}$ International Joint Conference on Artificial Intelligence. These interactions highlighted some of their expectations, concerns and dislikes regarding systems such as ours, its possible applications and counterfactual explanations in general. Here, we synthesise all of the lessons that we have learnt into four themes comprising safety of an AI system: explanations, model debugging, fairness, and security/privacy.

Our goal here is not to compare one explainability method against another to find the one that maximises safety of the underlying AI system by identifying and mitigating its harmful behaviour. Instead, we take a step back and reconsider how explainability may affect a predictive system. Our approach is motivated by an observed lack of evaluation and analysis of design choices that are made when proposing new explainability approaches. In our experience, it is uncommon among contributions in this space to provide a consideration of privacy, safety, security and adverse effects of an explanation.

We start by introducing the system that we used during all the demonstrations in Section . Then, after introducing the four aforementioned explainability themes in Sections –, we review relevant literature (Section ) followed by conclusions and future work (Section ).

## Class-contrastive Counterfactuals

Class-contrastive counterfactual explanations are well-suited for explaining tabular data given their accessibility and transparency. They are "user-friendly" and compliant (Wachter, Mittelstadt, and Russell 2018) with the European Union's General Data Protection Regulation[4] that came into force in May 2018 requiring organisations that use algorithmic decision making to provide their explanations on the client's request (Goodman and Flaxman 2017). Counterfactuals are furthermore versatile enough to explain predictions, express their fairness and help debug the underlying models.

Therefore, in our research, we decided to use them to explain logical AI systems, decision trees in particular. Since we have access to the model's internal structure our method guarantees to produce all the possible counterfactuals for a particular data point. This, in turn, allows us to focus on the counterfactuals themselves rather than their generation process and validity. This means that we have full control over their generation, hence we can tune it to the research question that we are posing. Whenever possible, we annotate every feature in a human understandable way (natural language description), we indicate which features are actionable from a user's perspective (e.g. age vs. the number of credit cards) and we make a note of protected attributes in the feature space (e.g. gender or race).

All these meta-data allows us to use our counterfactual generation process within a conversational system that can explain automated predictions, check their fairness and provide actionable suggestions to the user in a casual conversation. Our system has two interaction modes: a text-based chat served to the user as a web page and a voice-driven interaction facilitated by an off-the-shelf virtual personal assistant device such as Amazon Alexa or Google Home. Given our meta-annotations and the natural language interface the user can ask for: the shortest explanation, an explanation (not) using a particular feature as a foil, an explanation that is actionable and, finally, whether the prediction is fair. All of the examples presented below were generated with this setup deployed on top of two decision trees, one trained for each data set introduced earlier.

## Explanatory Properties of Counterfactuals

Counterfactual explanations have many advantages: they are short and easy to understand; they can be actionable; and they are interactive and delivered in a natural language, hence their complexity can be tuned to the recipient's requirements. Here are two examples:

[4] `https://publications.europa.eu/s/inbX`.

**Example 1.** Some explanations for the FICO explainable ML challenge data set:

> The prediction is **Bad**. It would be **Good** had the **Number of Satisfactory Trades** been *less or equal to* **10** instead of being **20**.
>
> The prediction is **Bad**. It would be **Good** had the **Number of Trades that has Ever been up to 60 Days Overdue and are marked as Derogatory in the Public Record** been *equal to* **0** instead of being **2**.

Counterfactual explanations also have some limitations that are not always explicitly recognised or stated. Firstly, they can be actionable but they are not causal. In their pure form they are local and their insights must not be generalised (which humans tend to do (Rozenblit and Keil 2002)) to other data points. Furthermore, there are multiple open research questions that have to be addressed before deploying counterfactual explanations in a mission-critical setting.

One question is how to adjust the (language) complexity of an explanation based on the audience, for example the difference between one given to a loan applicant and one provided to a regulator. Moreover, since every counterfactual is specific to a particular data point, whenever possible, it should be accompanied by a context so that the recipient knows the limitations of its generalisation. Finally, there are usually multiple counterfactuals of the same quality – e.g., the same number of features that has been altered – and it remains an open question how to pick the right one(s) and whether all of them are of the same importance. If interpreted incorrectly, it is not impossible for counterfactuals to have an adverse effect on safety of an AI system by causing indirect harm to the involved individuals.

## Model Debugging with Counterfactuals

In addition to their explanatory properties, counterfactuals can help to identify bugs, errors and mistakes in the underlying predictive model. Since counterfactuals express logical conditions, they can be used to automatically identify cases where the model behaves not as intended.

**Example 2.** Some unexpected explanations for the UCI German Credit data set:

> Your loan application has been **declined**. If your **savings account** had had *more than* **100** pounds, you had **not** had a savings account or its status had been **unknown**, your loan application would be **accepted**.
>
> Your loan application has been **declined**. Assuming that you had **asked for** *less or equal* to **£663** or *between* **883** and **1285** pounds, instead of **836** pounds, your loan application would be **accepted**.

The first explanation states that it is better not to disclose information about one's savings account, or not to have one at all, than to have only a modest amount of savings. The second example highlights non-monotonicity of the model, which arises often with decision trees but might have been overlooked by the person training the model.

## Algorithmic Fairness with Counterfactuals

Counterfactual statements can also be used to uncover disparate treatment – a scenario where changing a value of a protected attribute affects the classification outcome. As shown in the example below, counterfactuals are helpful in auditing fairness of predictive algorithms, thereby improving their safety by highlighting their harmful behaviour.

**Example 3.** A biased explanation for the UCI German Credit data set:

> The outcome of your loan application would have **changed** had you been a **male** (single, married, separated, divorced or widowed) instead of being a **female** (married, separated or divorced).

## Security and Privacy of Counterfactuals

Finally, we consider the effect of counterfactual explainability on privacy and security aspects of the underlying AI system and its training data. Counterfactual explanations can be used by an adversary to game a model – recall Example 2 where not having a savings account at all is more beneficial than having a small amount of savings. Therefore an attempt to make an AI system safer – by making it transparent – may have the opposite effect. This observation indicates a close relationship between counterfactual explanations and adversarial attacks that needs to be addressed before deploying explainable systems.

More generally, explanations reveal information about the underlying model and its training data. This issue is more prominent for counterfactual explanations than for other types of explanation since they leak information about decision boundaries. For example, for logical models these are the exact feature splits, for $k$-nearest neighbours these are training data points and for SVMs these could be the support vectors. This observation leads to the question: how many explanations of how many data points does it take to gather enough information to steal or game a model or its part?

This is further compounded as there are usually multiple counterfactual explanations of different length for a single data point, and revealing all of them can facilitate easier model stealing. Also, long counterfactuals where the foil is a conjunction of multiple logical conditions can reveal a big chunk of a model with just one explanation.

**Example 4.** An explanation for the FICO explainable ML challenge data set that raises security and privacy concerns:

> The prediction would have been **Good**, instead of **Bad**, had:
>
> - the **Number of Instalment Trades With Balance** been *less than* **3** instead of **3**,
> - the **Number of Revolving Trades With Balance** been *less than* **3** instead of **5**,
> - the **Number of Trades that has Ever been up to 60 Days Overdue and are marked as Derogatory in the Public Record** been *equal to* **0** instead of **2**, and
> - the **Number of Loans taken in the Last 12 Months** been *less or equal to* **2** instead of **5**.

Therefore, we believe that alongside every explainability approach the author should provide a critical evaluation of its privacy and security implications and a discussion about mitigating these factors to benefit the overall safety of an AI system.

## Related Work

The most prominent work discussing contrastive and counterfactual explanations from both computer and social sciences perspectives is Miller (2019). The author concludes that contrastive and counterfactual statements are the most natural explainability approach for humans interacting with intelligent systems. Wachter, Mittelstadt, and Russell (2018) present an optimisation approach to generating counterfactual statements for differentiable predictive models such as neural networks, support vector machines and regressors. Tolomei et al. (2017) also use a particular type of counterfactual statements – involving features, which when tweaked transform a true negative instance into a one that a model classifies as positive – generated for tree ensembles and used for improving on-line advertisements. Kusner et al. (2017) show that, in addition to their explanatory powers, counterfactual statements can also be used as a tool to audit fairness of AI agents. Gunning (2017) argued that explainability is an important step towards achieving safe artificial intelligence systems.

## Conclusions and Future Work

In this paper we investigated challenges and opportunities of counterfactual explainability in AI systems. We showed their advantages – interpretability, fairness and model debugging – and presented open research questions in this space. We also discussed security challenges that they pose focusing on model and training data stealing and gaming. All our observations are based on experience and are supported with examples arising in real data from the financial domain.

Our experiments have shown that when improving safety of an AI system by making it more transparent and explainable, one can unintentionally make it less secure and leak private data. Examples provided in this paper clearly show that security and privacy of a predictive model and its training data can be compromised when the influence of their explanations on the overall safety of the AI system is not assessed in the first place. All in all, this demonstrates that improving safety of an AI system is challenging and may have unexpected consequences.

Given the transparency of our counterfactual generation approach our future work will focus on the security of predictive logical models when explaining them with counterfactual statements. In particular, we are interested in identifying the least number of counterfactual explanations that are necessary to reverse-engineer or game a predictive model. Such research can be of importance for domains that need to find a balance between security and transparency of their AI systems.

## References


Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; Layton, R.; VanderPlas, J.; Joly, A.; Holt, B.; and Varoquaux, G. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 108–122.

Buolamwini, J., and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A., and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91. New York, NY, USA: PMLR.

Goodman, B., and Flaxman, S. 2017. European union regulations on algorithmic decision-making and a "right to explanation".

Gunning, D. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*.

Kulesza, T.; Burnett, M.; Wong, W.-K.; and Stumpf, S. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, 126–137. ACM.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 4069–4079.

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

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rozenblit, L., and Keil, F. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science* 26(5):521–562.

Tolomei, G.; Silvestri, F.; Haines, A.; and Lalmas, M. 2017. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 465–474. ACM.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2018. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology, 31 (2)*.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, 1171–1180. International World Wide Web Conferences Steering Committee.