Do Joint Energy-Based Models Produce More Plausible Counterfactual Explanations?

Giacomo Pezzali[1]
Supervisors: Cynthia C. S. Liem[1]**, Patrick Altmeyer**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Giacomo Pezzali
Final project course: CSE3000 Research Project
Thesis committee: Cynthia C. S. Liem, Patrick Altmeyer, Bernd Dudzik

Abstract

Counterfactual explanations (CEs) can be used to gain useful insights into the behaviour of opaque classification models, allowing users to make an informed decision when trusting such systems.

Assuming the CEs of a model are faithful (they well represent the inner workings of the model), an explainable model generates plausible CEs (i.e. CEs fitting the real-world distribution of the data). This raises the question of whether classifiers explicitly designed to model the distribution of the data, such as energy-based models, are inherently more explainable.

This work focuses on the evaluation of joint energy-based models (JEMs) in combination with the Energy-Constrained Conformal Counterfactuals (ECCCo) generator, with the goal of identifying if the generative capability of a model influences its explainability. Since ECCCo has been designed specifically to generate more faithful CEs, it makes it possible to use the CEs plausibility as a proxy of the model explainability.

Two experiments have been performed to evaluate the effect of variations of generative capability within the same JEM architecture and the difference between JEMs and classically trained classifiers. Despite the experiments not having established a clear correlation between generative capability and explainability of a model, various research avenues are still open to explore in future works.

1 Introduction

As AI systems become more ubiquitous in our daily life, the need to understand how these models take the decisions we ask them to take becomes more apparent. A responsible adoption of these technologies requires that we not only trust the developers of these tools, but the tools themselves [1]. The field of Explainable Artificial Intelligence (XAI) focuses on the study of tools that can be used to gain insight into Machine Learning models, with the aim of making the behaviour of these models more transparent and understandable for humans.

Counterfactual explanations (CEs) [2] are an intuitive tool to gain insights into a black-box model's behaviour. They are particularly useful in the context of algorithmic recourse (AR) [3], when an opaque classifier is used to make decisions that affect important aspects of human life and the people affected need a reason of why they were classified in a undesirable class and want an actionable plan to change this outcome. A CE is a set of changes that, when applied to a given input called "factual", result in the factual being classified as the desirable class. In general, the same factual can have infinitely many CEs.

A common example would be a classifier tasked with accepting or rejecting loan applications. If, for example, Alice's loan application gets rejected, a CE for her situation could be that she needs to earn 10% more, or that she needs to stay in her current job for at least another year.

Fundamental to the perceived trustworthiness of a system, is the plausibility of a CE, defined by how much the new input looks like an outcome of the same probability distribution as the observed points of the target class. In our example, if Alice was told that she needs to keep her job for another year and she knows that her colleague Beth just got her loan approved after working in the same place for one year and three months more than her, she would consider this CE plausible. But if she was told instead that she needs to earn the entire GDP of The Netherlands for her application to be accepted, she would certainly trust the system a lot less.

Another important characteristic of a CE is its faithfulness: how closely the CE represents the internal "understanding" of the classifier. Faithfulness and plausibility are closely related, since faithful CEs generated for a model that has correctly internalised the training data will also be plausible and vice versa [4]. For this reason, a model is considered "explainable" if its faithful CEs are also plausible.

Altmeyer, Farmanbar, van Deursen, *et al.* [4] have proposed to use "Energy-Constrained Conformal Counterfactuals" (ECCCo) as a novel way of generating faithful CEs, taking ideas from joint energy-based modelling [5] and conformal predictions [6].

Since ECCCo makes use of the energy-based generative capabilities of the model to penalise unfaithful CEs, it can be reasonably hypothesised that models specifically trained to focus on such generative capabilities would be more explainable. Joint Energy-Based Models (JEM), proposed by Grathwohl, Wang, Jacobsen, *et al.* [5], do exactly that, as they train a classifier architecture with the double objective of acting as a classifier as well as a generator.

The aim of this work is to evaluate whether the generative capability of a model affects the plausibility of the CEs produced with the ECCCo technique. With this in mind, two questions need to be answered:

1. Does the generative capability of a joint-energy model affect the plausibility of the generated CEs?
2. Given the same architecture, are the CEs generated from a joint-energy model more plausible than the ones generated on a classical model?

To answer these questions for a general neural network, different architectures will be considered for a number of common datasets, trained both classically and as joint-energy models and used to generate CEs.

The rest of this paper will first provide a more rigorous description of the ECCCo generator and joint-energy models (section 2), followed by a description of the experiments performed to answer the research questions (section 3). The results of these experiments will be then presented (section 4) and commented (section 5). Subsequently, the author will reflect on the ethical implications of this research (section 6), on its limitations and on the questions that it leaves unanswered (section 7). Finally, all these elements will be put together to answer this work's research questions (section 8).

2 Background

2.1 Counterfactual generation

In the process of generating counterfactual explanations, a few different aspects need to be considered. A good CE needs to be:

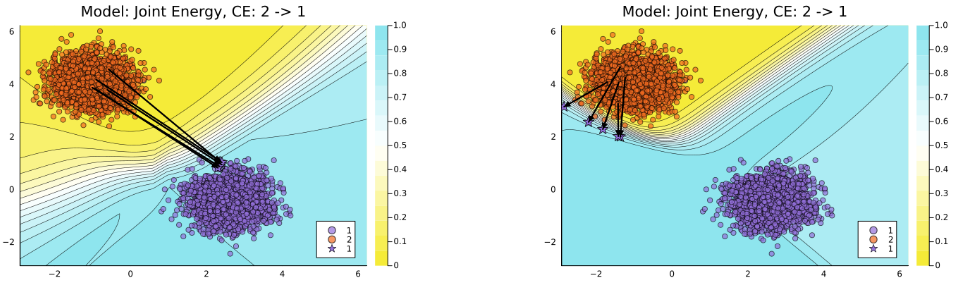
- **Valid**: a CE is valid if the suggested change would actually result in the desired classification.
- **Plausible**: a CE is plausible if it matches the distribution of real-world data [7].
- **Faithful**: a CE is faithful if it reflects the internal "understanding" of the model [4]. For example, if a loan-requests classifier has learned that tall applicants are usually granted loans, a faithful counterfactual explanation would suggest to an applicant to become taller.
- **Actionable**: a CE is actionable if it suggests changes that can actually be applied [8]. The CE suggesting applicants to become taller would not be actionable, while one suggesting to increase by 5% their annual income would be.
- **Close**: a CE is closer the smaller the changes that it suggests are [9]. In the example above, a 1% increase in the annual income is closer than a 50% increase.

To better illustrate the difference between faithfulness and plausibility, Figure 1a shows counterfactual explanations that are both plausible (i.e. indistinguishable from the training data) and faithful. Additionally, Figure 1b shows CEs that are faithful, being representative of the internal understanding of the model, but not plausible, as they look more like outliers of class 2 than members of class 1.

There are various techniques to generate counterfactual explanations, depending on whether the model is differentiable or not [10]. This work focuses on the Energy-Constrained Conformal Counterfactuals (ECCCo) generator, proposed by Altmeyer, Farmanbar, van Deursen, *et al.* [4], in particular, on the ECCo variant (i.e. without the Conformal Prediction component).

The ECCo generator is based on gradient descent, with the following objective function:

$$\min_{x' \in \mathcal{X}} \left\{ L_{\text{clf}}\left(f(x'); M_\theta, c_T\right) + \lambda_1 \cdot \text{cost}(f(x')) + \lambda_2 \cdot \varepsilon_\theta(f(x') \mid c_T) \right\}$$



(a) Example of CEs both plausible and faithful.

(b) Example of CEs faithful but not plausible.

Figure 1: Examples of CEs targeting class 1 of the Linearly Separable dataset on two different classifiers.

Where x' is the counterfactual explanation, $\mathcal{X}$ is the counterfactual domain, $L_{\text{clf}(\cdot)}$ is a standard classification loss, $f(\cdot)$ maps the counterfactual state space into the domain state space (usually $f(\cdot)$ is the identity function), M_θ is the model, c_T is the target class, λ_1 is the relative penalty for distance, $\text{cost}(\cdot)$ is the distance from the factual, λ_2 is the relative penalty for unfaithfulness and $\varepsilon_\theta(\cdot \mid c_T)$ is the energy conditioned to the target class (the negative of the logit corresponding to c_T) [4].

The CE desiderata are achieved in different ways during the generation process: actionability is ensured by limiting the search space $\mathcal{X}$, using domain specific knowledge of the input parameters; closeness and faithfulness are achieved by the λ_1 and λ_2 penalties respectively and validity is given by L_{clf}. However, practical limitations such as the maximum number of iterations in the process or too restrictive actionability constraints, might lead to invalid CEs.

Plausibility can be measured knowing the training data of the model. The metric used by Altmeyer, Farmanbar, van Deursen, *et al.* [4] and in this work is its inverse, implausibility, which is defined as follows:

$$\text{impl}(x', \mathcal{X}_{c_T}) = \frac{1}{|\mathcal{X}_{c_T}|} \sum_{x \in \mathcal{X}_{c_T}} \text{dist}(x, x')$$

where x' is the counterfactual explanation, $\mathcal{X}_{c_T}$ is the subset of points from the training set originally labelled as the target class c_T and $\text{dist}(\cdot, \cdot)$ is the euclidean distance. $\mathcal{X}_{c_T}$ is limited to the 100 nearest neighbours of x'.

As mentioned in section 1, a model is considered explainable if its faithful CEs are also plausible, therefore a lower implausibility is desirable.

2.2 Joint energy-based models

Joint Energy-based Models (JEMs) are a form of hybrid models, capable of acting both as classifier and as generator, introduced by Grathwohl, Wang, Jacobsen, *et al.* [5]. The specifics of the implementation of these models are beyond the scope of this work, however, the general concept is that a classifier architecture (such as a multi-layer perceptron) can be trained to output the energy distribution of the dataset, meaning that, given an input x, the logit of class c_i will take the value $E(x \wedge c_i)$. Energy is defined as per the Boltzmann distribution: $p(j) = \frac{e^{-E(j)}}{\int_{\mathcal{K}} e^{-E(k)}\,\mathrm{d}k}$, where $\mathcal{K}$ is the input domain. It follows that $p(x \wedge c_i)$ can be computed for every class and used to evaluate $p(c_i \mid x)$ (classification task) and $p(x \mid c_i)$ (generative task) [5].

During the training of a JEM, both the accuracy (quality of the classification task) and the generative loss (quality of the generative task) of the model are monitored and optimised. The generative loss is defined as the energy of the training data minus the energy of samples generated by the model.[1] Generative loss will be used in this paper to evaluate a model's generative capability; a lower value of generative loss indicates a more effective generator.

[1] `https://github.com/JuliaTrustworthyAI/JointEnergyModels.jl/blob/38fd7d7ba4b34a06d12bb2d3df0cdd89ffabd928/src/model.jl#L61`

3 Methodology

Two experiments were conducted to investigate the relation between energy-based training and plausibility of counterfactual explanations: the first to compare joint-energy classifiers with different generative losses, the second to compare the classically trained classifiers and joint-energy classifiers. The experiments were ran in Julia,[2] using modules developed by the "Trustworthy Artificial Intelligence in Julia" (Taija) group.[3] In particular: `TaijaData` was used to load the relevant datasets, `TaijaParallel` to speed up the generation of counterfactual explanations, `CounterfactualExplanations` to generate and manipulate the counterfactual explanations and `JointEnergyModels` to train the joint energy-based models.

For both experiments, multiple pairs of dataset and neural network architecture were defined. The datasets chosen are either artificially generated or commonly used in the literature [10]:

- **Circles**: an artificial dataset consisting of a central blob of class 1 and a corona surrounding it of class 2.
- **Linearly Separable**: an artificial dataset consisting of two blobs, one for each class, that could be fully separated by a linear classifier.
- **Overlapping**: an artificial dataset consisting of two blobs, one for each class, that cannot be fully separated by a linear classifier.
- **MNIST**: an image dataset, consisting of 28x28 images of handwritten digits in grey-scale [11].
- **California Housing**: a financial dataset for house values, based on the 1990 California Census [12].
- **German Credit**: a financial dataset for credit risk [13].

Table 3 in the Appendix shows which architecture was applied for each dataset. Most architectures are simple multi-layer perceptrons (MLPs) defined for this study, while some are taken from the literature. In the rest of this paper, each pair is identified by the dataset used and, if applicable, the name of the author whose architecture was used.

3.1 Explainability evaluation

To evaluate the explainability of a model M, the same procedure was followed every time. For every combination of starting class c_S and target class c_T, a large amount of factuals is selected. Factuals are values from the training set that are labelled as c_S and classified as the same class by the model. From each factual, a CE with the given target class is generated, the implausibility of each CE is computed and the average of these values for all CEs is taken to obtain a single value $\widehat{\text{impl}}_M$.

To mitigate the effect of fortuitous selections of factuals, the same procedure was repeated multiple times, using the average $\hat{\mu}_{\text{impl}_M}$ and standard deviation $\hat{\sigma}_{\text{impl}_M}$ of each estimated $\widehat{\text{impl}}_M$ as a way to compare different models.

In this framework, a model is more explainable the lower its implausibility score $\hat{\mu}_{\text{impl}_M}$ is.

Ideally, the number of CEs with starting class c_S and target class c_T is the same for each combination of c_S and c_T (excluding the case in which $c_S = c_T$). An approximation of this even distribution was obtained selecting a number of factuals equal to $\lceil n/(c \cdot (c-1)) \rceil$ (where n is the desired number of factuals and c is the number of classes in the dataset) for each valid combination of c_S and c_T. However, the amount of possible factuals with a given starting class was not always as high as this formula required, due to additional constraints given by the experiment, therefore the final number of CEs computed for each model did not always amount to the desired n.

Counterfactual explanations were generated using the ECCCo technique without the conformal prediction component, since the full implementation was not available at the time of running the experiments. The penalties for the gradient descent procedure were kept at the default values, as the time available did not allow for a proper tuning of these parameters.

[2] `https://julialang.org/`

[3] `https://github.com/JuliaTrustworthyAI`

3.2 Intra-model experiment

To evaluate the effect of different generative losses, the same model was trained using the joint-energy technique multiple times on the same data, using a different random initialisation each time. For each trained model, its generative loss and its average implausibility were noted. The same procedure was repeated for each dataset-architecture pair.

The correlation coefficient between these values was used to estimate if a relation exists between the generative capabilities of a model and its explainability. A correlation coefficient close to 1 would indicate that models with strong generative capabilities (low generative loss) are generally more explainable (low implausibility score), as hypothesised in this work.

3.3 Training-based experiment

For the training-based experiment, each pair of dataset and architecture was trained twice, once in the classical way and once using joint-energy training. By "classical training", it is intended the state of the art use of backpropagation to estimate the weights of the neurons' connections in a neural network given a training dataset.

To compare the two trained models, the same set of factuals was selected from the training dataset, making sure that both models correctly classified each of them.

The goal of this experiment is to use the classical training as the baseline and to consider a joint-energy model significantly different from the baseline if its implausibility score is more than $\hat{\sigma}_{\mathrm{impl}_{\mathrm{CM}}}$ away from the classical model's implausibility score. This experiment's hypothesis is that

$$\hat{\mu}_{\mathrm{impl}_{\mathrm{CM}}} - \hat{\mu}_{\mathrm{impl}_{\mathrm{JEM}}} \geq \hat{\sigma}_{\mathrm{impl}_{\mathrm{CM}}}$$

for all dataset-architecture pairs.

4 Experimental setup and results

For both experiments, the generation of CEs was performed using the `ECCoGenerator` available in the `CounterfactualExplanations.jl` module with $\lambda_1 = 0.1$ and $\lambda_2 = 0.5$. The conditions for convergence were set at 0.01 gradient tolerance and default values for every other condition (0.5 decision threshold, 100 max iterations and 0.75 min success rate).

The search space was limited for each dimension of the input to double the range of the training data, keeping the same middle point: for a dimension d whose training data-points spanned $[m_d, M_d]$ the CE search was limited to $\left[m_d - \frac{M_d - m_d}{2}, M_d + \frac{M_d - m_d}{2}\right]$.

As described in section 3, $\widehat{\mathrm{impl}}_M$ was estimated 5 times attempting to select 1000 factuals each time, to a total of at most 5000 CEs per model. In the intra-model experiment, each architecture was trained and evaluated 10 times. Table 4 and Table 5 in the Appendix show how many CEs were computed for each dataset-architecture pair and how many of those CEs were valid. They also provide information about the accuracy of the models.

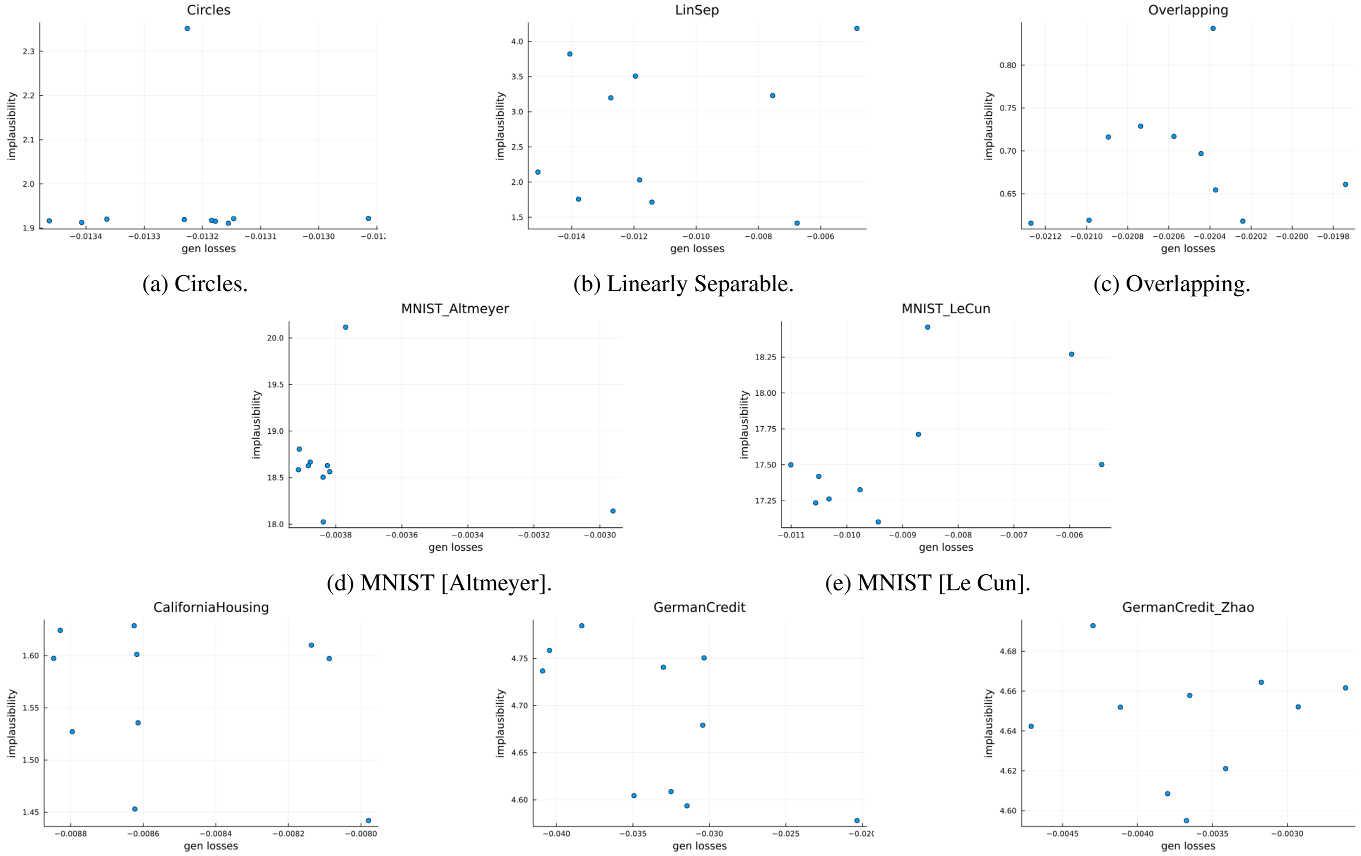
4.1 Results

The results of the intra-model experiment can be found in Table 1, showing the correlation between generative loss and implausibility (computed only on the valid CEs) for each dataset-architecture pair. Additionally, Figure 2 provides a visual representation of the same results, with plots showing the generative loss of a model against the implausibility of the CEs generated.

Table 2 shows the results of the training-based experiment for the different datasets. Each dataset-architecture pair presents the implausibility computed using all available CEs as well as the implausibility computed using only the valid ones for both classical and energy based models. Finally, the plot of model's generative loss against valid CEs' implausibility can be found in Figure 3.

Table 1: Intra-model correlation between generative loss and implausibility for each dataset-architecture pair. The author of the architecture is mentioned for non-original architectures. Values above 0.5 and below -0.5 are in bold face.

Dataset	**Correlation**
Circles	0.0117
Linearly Separable	0.2056
Overlapping	0.1506
MNIST [Altmeyer]	-0.2536
MNIST [Le Cun]	**0.5142**
California Housing	-0.2460
German Credit	**-0.6261**
German Credit [Zhao]	0.0078



(a) Circles.

(b) Linearly Separable.

(c) Overlapping.

(d) MNIST [Altmeyer].

(e) MNIST [Le Cun].

(f) California Housing.

(g) German Credit.

(h) German Credit [Zhao].

Figure 2: Intra-model generative loss vs. valid CE implausibility for all dataset-architecture pairs.

Table 2: Results of the training-based experiment for all dataset-architecture pairs. The author of the architecture is mentioned for non-original architectures. The lower implausibility of each column is in bold face. Values for joint-energy models below one (*) or two (**) and above one (†) or two (††) standard deviations from baseline are marked.

Training	**Circles**		**Linearly Separable**		**Overlapping**	
	Implausibility (all CEs)	Implausibility (valid CEs)	Implausibility (all CEs)	Implausibility (valid CEs)	Implausibility (all CEs)	Implausibility (valid CEs)
Classical	**0.38 ± 0.25**	**0.33 ± 0.30**	4.24 ± 0.51	2.37 ± 0.03	**0.91 ± 0.20**	**0.85 ± 0.26**
Joint-Energy	3.16 ± 0.56 ††	1.91 ± 1.87 ††	**3.17 ± 0.78** **	**2.13 ± 0.28** **	1.56 ± 0.86 ††	0.91 ± 0.20

Training	**MNIST [Altmeyer]**		**MNIST [Le Cun]**	
	Implausibility (all CEs)	Implausibility (valid CEs)	Implausibility (all CEs)	Implausibility (valid CEs)
Classical	18.80 ± 1.83	**16.50 ± 1.66**	18.02 ± 1.47	**17.07 ± 2.06**
Joint-Energy	**18.79 ± 1.83**	19.17 ± 2.49 †	**17.90 ± 1.51**	17.53 ± 1.82

Training	**California Housing**		**German Credit**		**German Credit [Zhao]**	
	Implausibility (all CEs)	Implausibility (valid CEs)	Implausibility (all CEs)	Implausibility (valid CEs)	Implausibility (all CEs)	Implausibility (valid CEs)
Classical	**1.62 ± 0.31**	1.62 ± 0.31	**4.75 ± 0.21**	**4.74 ± 0.43**	4.94 ± 0.07	4.90 ± 0.07
Joint-Energy	1.89 ± 0.10	**1.43 ± 0.43**	4.94 ± 0.05	4.75 ± 0.18	**4.81 ± 0.04** *	**4.72 ± 0.07** **

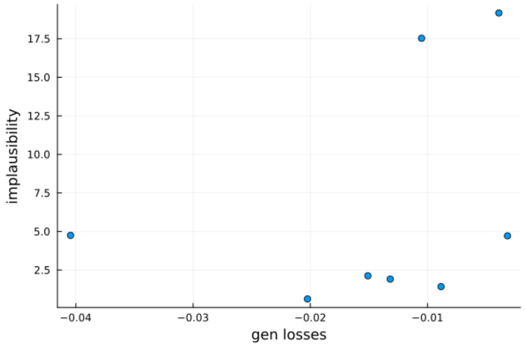


Figure 3: Generative loss vs. valid CE implausibility for all models of the training-based experiment.

5 Discussion

The relationship between generative loss of a model and its ability to generate plausible counterfactual explanations does not seem to be significant, as shown in Figure 3 and from the results of the intra-model experiment: for all dataset-architecture pairs, the correlation shown in Table 1 is fairly close to 0. Only MNIST [Le Cun] and German Credit got values above 0.3 (in absolute terms), but presenting still a very low coefficient showing no clear correlation between the two quantities.

For the training-based experiment, the different models seem to behave in a wide variety of ways: Circles and MNIST [Altmeyer] show a significant increase of implausibility in CEs generated on the joint-energy models; Overlapping, MNIST [Le Cun], California Housing and German Credit show no significant difference between the different training methods; and finally Linearly Separable and

German Credit [Zhao] show a consistent decrease of implausibility for JEMs. These groupings do not seem to follow any particular pattern based on the model's accuracy nor on the percentage of valid CEs generated.

It is to be noted that certain models perform very differently depending on the "direction" of the counterfactual explanation being generated. The Circles models are a good example of this phenomenon, Figure 4 and Figure 5 show clearly how in the classical model, the implausibility of CEs targeting class 2 is much higher than the ones targeting class 1. While for the energy-based model the opposite applies, with CEs targeting class 2 being extremely implausible (albeit very faithful). Furthermore, the model trained classically on the Linearly Separable dataset did not produce any valid CE targeting class 1 (Figure 6), probably due to the limited number of steps allowed to the ECCCo generator and the unlucky shape of the model's decision boundary.

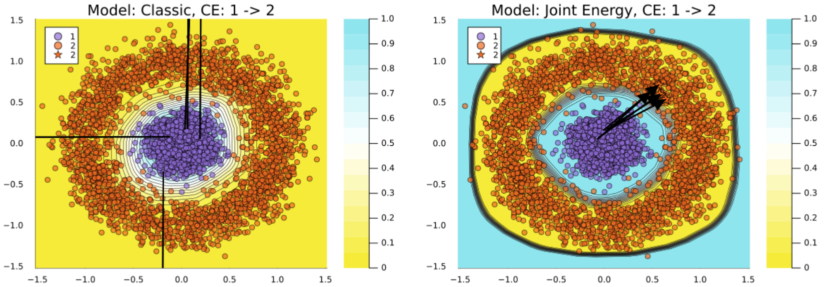


Figure 4: Examples of CEs targeting class 2 of the Circle dataset on the models trained classically (left) and with energy-based training (right).

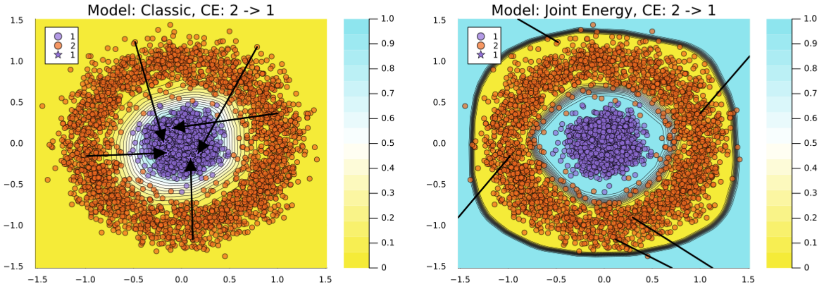


Figure 5: Examples of CEs targeting class 1 of the Circle dataset on the models trained classically (left) and with energy-based training (right).

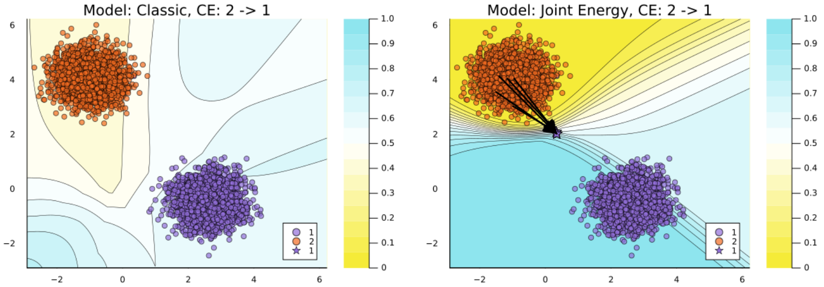


Figure 6: Examples of CEs targeting class 1 of the Linearly Separable dataset on the models trained classically (left) and with energy-based training (right).

6 Responsible research

From an ethical perspective, this work could have run into two main issues: lack of reproducibility and confirmation bias of the author towards the "desirable" outcome of a strong positive influence of a model's generative capability on the counterfactual explanations plausibility.

To ensure a high degree of reproducibility, a series of precautions were taken: all the code is publicly available on GitHub [4] and only free and/or open source technologies were used to execute the experiments.
The scripts for both experiments reset the random seed when run, therefore the same models will be trained and the same factuals and counterfactual explanations should be obtained when re-running each of them. However, the results for such a clean re-run of the experiments would not be exactly the same as the one reported in this paper: in order to reduce the time needed to run the experiments, some of the models were pre-trained, therefore the number of draws from the RNG buffer was altered. The pre-trained models are available in the same repository as a release artefact and they should still give results comparable to the ones reported here.
Additionally, each experiment activates a Julia environment that sets the specific version of the language and modules to be used, therefore even if substantial changes happen in the future, the code will remain true to itself.

Aside the specific repeatability of each individual run, the experiments are designed to account for the randomness of the counterfactual generation: a high number of CEs is generated for each model multiple times (randomly selecting factuals each time) to balance possible "lucky" runs. This setup makes the results more stable and helps balancing the possibility of unconsciously choosing to only publish the results a particularly successful run.

A general ethical consideration about the use of machine learning models, is that biases present in the training data and learned by the model could be exacerbated. In this study in particular, the faithfulness and the plausibility of a CE are kept in high regard, but both of these qualities are heavily affected by the distribution of the training data and would therefore reflect those biases. As an extreme example, a model that learns that male applicants have a higher chance of being hired by a company would produce a faithful counterfactual explanation for a female applicant who got rejected proposing the "become male" change. Such CE would be both faithful and plausible (given the historical bias in the training data), resulting in the model being considered more explainable. This means that users of machine learning tools should not make the mistake of confusing "explainable" with "trustworthy", since trustworthiness only comes from a critical analysis of the explainability of the model. The use of highly explainable models makes it easier to spot these biases and correct them or discard the models that perpetuate them. This work aims to provide ways to generate more explainable models, but using these models ethically is a burden that falls on us all.

7 Limitations and future work

A series of factors have contributed to the inconclusiveness of the results. In particular, this work presents limitations in the experimental setup, in the model training and in the CE generation.

The main issue with the experimental setup stems from the attempt of generalising the results to a wide spectrum of models: the number of possible neural network architectures is infinite, as it is the range of their applications. A more focused approach could be beneficial, for example choosing only two possible architectures, one with many nodes and layers and another very simple, and limiting the datasets to binary classification problems.

The second set of issues lies with the training of the joint energy-based models. The approach taken was to train them all in similar conditions, changing only the weight initialisation, while two things could have been noted instead:

1. For the intra-model experiment, different iterations could have been trained prioritising more or less the generative task of the model, in order to have a wider spread of generative loss values and possibly more meaningful correlations.
2. In general, JEM's training is particularly unstable, and it is therefore quite hard to obtain faithful models without supervision. Grathwohl, Wang, Jacobsen, *et al.* [5] suggest ways to improve on this issue, but no such precaution was taken in this work due to the limited amount of time and computational power available.

Finally, the last category of limitations lies with the generation of counterfactual explanations: when using the `ECCoGenerator`, the penalties have been kept constant for all the datasets examined, but

[4] `https://github.com/JuliaTrustworthyAI/what-makes-models-explainable`

it is now clear how different distributions have different priorities. For example the MNIST dataset would benefit from a less restrictive penalty on closeness, since switching from one digit to another requires significant changes to many pixels, while the Circle dataset has the opposite problem, with many CEs reaching the extremes of the search-space. In general, taking the time to tune the hyperparameters of the generator, both in terms of penalties and stopping conditions, could have helped obtain more conclusive results and avoid some of the issues discussed in section 5 too.

In light of these limitations, future works expanding on this research can take a variety of directions:

1. Explore more accurately the intra-model explainability, systematically changing the balance of generation and classification task when training the joint-energy models.
2. Focus on the stabilisation of the JEM training process, in order to evaluate the explainability of models that actually present the advertised generative capability.
3. Work on the generalising the results, taking a more rigorous approach to the definition of the architectures to use and evaluating them on the same dataset.

8 Conclusions

This work has tried to answer the questions of whether joint-energy models become more explainable as their generative capability increases and if joint-energy models are in general more explainable than classically trained neural network models.

To provide answers to these questions, two experiments have been designed and applied on various commonly used dataset. The first training a joint-energy model multiple times and comparing the obtained generative capacities with the corresponding explainabilities. The second comparing the explainability of classically trained models with the one of joint-energy models, given the same underlying architecture. The plausibility of counterfactual explanations generated using ECCCo was used as a proxy for the explainability of a model.

In the conditions under which the experiments have been conducted, no relevant influence was found of the generative capability on the explainability of a joint-energy model. Furthermore, the use of such technique was shown to produce more explainable models only in certain cases, while other cases were better served by classical training. No pattern emerged based on characteristics of the datasets or the models used.

The investigation on joint-energy models' explainability in the context of counterfactual explanations, however, has not reached a dead end. Changing the experimental conditions could yield interesting results, possibly even subvert the findings of this work. In particular, repeating the first experiment while increasing the importance of the generative training objective of the models could provide a wider overlook of the influence of a model's generative capability on its explainability. Furthermore, putting additional effort into counterbalancing the training instability of joint-energy models could result in their explainability being more clearly distinct from classically trained models.

References

[1] D. Petkovic, "It is Not "Accuracy vs. Explainability" - We Need Both for Trustworthy AI Systems," *IEEE Transactions on Technology and Society*, vol. 4, no. 1, pp. 46–53, 2023. DOI: 10.1109/TTS.2023.3239921.

[2] P. Altmeyer, A. van Deursen, and C. C. S. Liem, "Explaining Black-Box Models through Counterfactuals," *Proceedings of the JuliaCon Conferences*, vol. 1, no. 1, p. 130, 2023. DOI: 10.21105/jcon.00130. [Online]. Available: https://doi.org/10.21105/jcon.00130.

[3] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera, "A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects," 2021. arXiv: 2010.04050.

[4] P. Altmeyer, M. Farmanbar, A. van Deursen, and C. Liem, "Faithful Model Explanations through Energy-Constrained Conformal Counterfactuals," English, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 10, pp. 10 829–10 837, 2024, ISSN: 2159-5399. DOI: 10.1609/aaai.v38i10.28956.

[5] W. Grathwohl, K.-C. Wang, J.-H. Jacobsen, D. Duvenaud, M. Norouzi, and K. Swersky, “Your classifier is secretly an energy based model and you should treat it like one,” in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=Hkxzx0NtDB.

[6] A. N. Angelopoulos and S. Bates, *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*, 2022. arXiv: 2107.07511.

[7] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh, “Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems,” 2019. arXiv: 1907.09615.

[8] B. Ustun, A. Spangher, and Y. Liu, “Actionable Recourse in Linear Classification,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, ser. FAT* ’19, ACM, Jan. 2019. DOI: 10.1145/3287560.3287566. [Online]. Available: http://dx.doi.org/10.1145/3287560.3287566.

[9] S. Wachter, B. D. Mittelstadt, and C. Russell, “Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR,” *CoRR*, vol. abs/1711.00399, 2017. arXiv: 1711.00399. [Online]. Available: http://arxiv.org/abs/1711.00399.

[10] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, and C. Shah, “Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review,” 2022. arXiv: 2010.10596.

[11] Y. LeCun, C. Cortse, and C. Burges. “THE MNIST DATABASE of handwritten digits.” (1998), [Online]. Available: http://yann.lecun.com/exdb/mnist/.

[12] R. Kelley Pace and R. Barry, “Sparse spatial autoregressions,” *Statistics & Probability Letters*, vol. 33, no. 3, pp. 291–297, 1997, ISSN: 0167-7152. DOI: https://doi.org/10.1016/S0167-7152(96)00140-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016771529600140X.

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

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998. DOI: 10.1109/5.726791.

[15] Z. Zhao, S. Xu, B. H. Kang, M. M. J. Kabir, Y. Liu, and R. Wasinger, “Investigation and improvement of multi-layer perceptron neural networks for credit scoring,” *Expert Systems with Applications*, vol. 42, no. 7, pp. 3508–3516, 2015, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2014.12.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417414007726.

Appendix

Table 3: Neural Network architectures used for each dataset in the experiments. If an architecture is taken from a previous work, a specifier with the name of the first author and a reference are added.

Dataset	**Architecture**
Circles	Input: 2 nodes, Activation: relu Dense: 32 nodes, Activation: relu Dense: 32 nodes, Activation: relu Dense: 32 nodes, Activation: identity Output: 2 nodes, Activation: identity
Linearly Separable	Input: 2 nodes, Activation: tanh Dense: 4 nodes, Activation: relu Dense: 4 nodes, Activation: identity Output: 2 nodes, Activation: identity
Overlapping	Input: 2 nodes, Activation: relu Dense: 16 nodes, Activation: relu Dense: 16 nodes, Activation: identity Output: 2 nodes, Activation: identity
MNIST [Altmeyer] [2]	Input: 784 nodes, Activation: relu Dense: 32 nodes, Activation: tanh Output: 10 nodes, Activation: identity
MNIST [Lecun] [14]	Input: 28x28 nodes, 1 channel, Activation: identity Convolution: 5x5 kernel, 1-padding, 6 channels, Activation: relu Pooling: 2x2 window Convolution: 5x5 kernel, 1-padding, 16 channels, Activation: relu Pooling: 2x2 window Flatten: 7x7x16 to 784 nodes Dense: 400 nodes, Activation: relu Dense: 120 nodes, Activation: relu Dense: 84 nodes, Activation: relu Output: 10 nodes, Activation: relu
California Housing	Input: 8 nodes, Activation: relu Dense: 32 nodes, Activation: relu Dense: 128 nodes, Activation: relu Dense: 32 nodes, Activation: relu Output: 2 nodes, Activation: identity
German Credit	Input: 10 nodes, Activation: relu Dense: 128 nodes, Activation: relu Dense: 128 nodes, Activation: relu Output: 2 nodes, Activation: identity
German Credit [Zhao] [15]	Input: 10 nodes, Activation: tanh Dense: 10 nodes, Activation: tanh Output: 2 nodes, Activation: identity

Table 4: Information about the models for the intra-model experiment: accuracy mean and standard deviation, number of generated CEs and number of valid CEs.

Dataset	**Accuracy**	**Total CEs**	**Valid CEs**
Circles	0.99 ± 0.00	50000	30301
Linearly Separable	0.96 ± 0.11	50000	32482
Overlapping	0.92 ± 0.00	50000	36332
MNIST [Altmeyer]	0.31 ± 0.07	47430	1899
MNIST [Le Cun]	0.85 ± 0.15	46440	21365
California Housing	0.87 ± 0.00	50000	25513
German Credit	0.92 ± 0.04	45765	30512
German Credit [Zhao]	0.77 ± 0.02	38670	31633

Table 5: Information about the models for the training based experiment: accuracy and proportion of valid CEs for each training technique.

Dataset	**Accuracy (classical)**	**Accuracy (JEM)**	**Total CEs**	**Valid CEs (classical)**	**Valid CEs (JEM)**
Circles	0.9976	0.9872	5000	3702	2715
Linearly Separable	0.9998	0.9998	5000	1078	2200
Overlapping	0.9198	0.9160	5000	4461	3007
MNIST [Altmeyer]	0.8141	0.2178	4140	120	131
MNIST [Le Cun]	0.6730	0.8928	3240	330	1431
California Housing	0.8372	0.8810	5000	5000	2503
German Credit	0.7860	0.9360	3530	671	4231
German Credit [Zhao]	0.6840	0.8030	3025	2330	2408