

Relevant Irrelevance: Generating Alterfactual Explanations for Image Classifiers

Silvan Mertes¹, Tobias Huber¹, Christina Karle¹, Katharina Weitz^{1,2},
Ruben Schlagowski¹, Cristina Conati³, Elisabeth André¹

¹University of Augsburg, Germany

²Fraunhofer HHI, Germany

³University of British Columbia, Canada

{silvan.mertes, tobias.huber, ruben.schlagowski, elisabeth.andre}@uni-a.de,
katharina.weitz@hhi.fraunhofer.de, conati@cs.ubc.ca

Abstract

In this paper, we demonstrate the feasibility of alterfactual explanations for black box image classifiers. Traditional explanation mechanisms from the field of Counterfactual Thinking are a widely-used paradigm for Explainable Artificial Intelligence (XAI), as they follow a natural way of reasoning that humans are familiar with. However, most common approaches from this field are based on communicating information about features or characteristics that are especially important for an AI’s decision. However, to fully understand a decision, not only knowledge about relevant features is needed, but the awareness of irrelevant information also highly contributes to the creation of a user’s mental model of an AI system. To this end, a novel approach for explaining AI systems called alterfactual explanations was recently proposed on a conceptual level. It is based on showing an alternative reality where irrelevant features of an AI’s input are altered. By doing so, the user directly sees which input data characteristics can change arbitrarily without influencing the AI’s decision. In this paper, we show for the first time that it is possible to apply this idea to black box models based on neural networks. To this end, we present a GAN-based approach to generate these alterfactual explanations for binary image classifiers. Further, we present a user study that gives interesting insights on how alterfactual explanations can complement counterfactual explanations.

1 Introduction

With the steady advance of Artificial Intelligence (AI), and the resulting introduction of AI-based applications into everyday life, more and more people are being directly confronted with decisions made by AI algorithms [Stone *et al.*, 2016]. As the field of AI advances, so does the need to make such decisions explainable and transparent. The development and evaluation of *Explainable AI* (XAI) methods is important not only to provide end users with explanations that increase acceptance and trust in AI-based methods, but also to empower

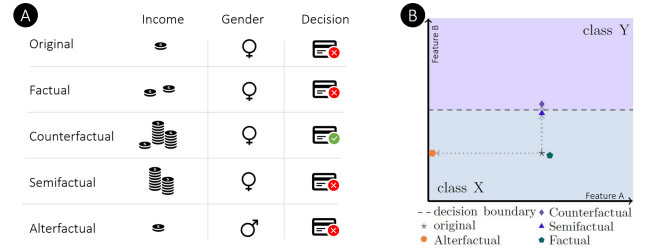


Figure 1: (A) Examples of a counterfactual and an alterfactual explanation. Input features to a fictional decision system to be explained are *Income* and *Gender*, whereas the former is relevant and the latter is irrelevant to the AI’s decision on whether a credit is given or not. (B) Conceptual comparison of factual, counterfactual, semifactual, and alterfactual explanations.

researchers and developers with insights to improve their algorithms.

The need for XAI methods has prompted the research community to develop a large variety of different approaches to unravel the black boxes of AI models. A considerable part of these approaches is based on telling the user of the XAI system in various ways *which* features of the input data are important for a decision (often called *Feature Attribution*) [Arieta *et al.*, 2020]. Other methods, which are close to human habits of explanation, are based on the paradigm of *Counterfactual Thinking* [Miller, 2019]. Procedures that follow this guiding principle try answering the question of *What if...?* by showing an alternative reality and the corresponding decision of the AI. Here, in contrast to feature attribution mechanisms, not only the importance of the various features is emphasized. Rather, it is conveyed, even if only indirectly, *why* features are relevant.

Prominent examples of these explanatory mechanisms are *Counterfactual Explanations* and *Semifactual Explanations* [Kenny and Keane, 2020]. Counterfactual explanations show a version of the input data that is altered just enough to change an AI’s decision. By doing so, the user is shown not only *which* features are relevant to the decision, but more importantly, *how* they would need to be changed to result in a different decision of the AI. Semifactual explanations follow a similar principle, but they modify the relevant features of the input data to an extent that the AI’s decision does not change.

All of these methods have in common that they focus on

the *important* features. However, we argue that awareness of irrelevant features can also contribute substantially to the complete understanding of a decision domain, as knowledge of the important features for the AI does not necessarily imply knowledge of the unimportant ones.

For example, consider an AI system that assesses a person’s creditworthiness based on various characteristics. If that system was completely fair, a counterfactual explanation might be of the form: *If your income was higher, you would be creditworthy*. However, this explanation does not exclude the possibility that your skin color also influenced the AI’s decision. It only shows that the income had a high impact on the AI. An explanation confined to the irrelevant features, on the other hand, might say *No matter what your skin color is, the decision would not change*. In this case, direct communication of irrelevant features ascertains fairness with regards to skin color. Conventional counterfactual thinking explanation paradigms do not provide this information directly. To address this issue, Mertes *et al.* [2022b] recently conceptually introduced the explanatory paradigm of *Alterfactual Explanations* that is meant to complement counterfactual explanations. This principle is based on showing the user of the XAI system an alternative reality that leads to the exact same decision of the AI, but where irrelevant features are altered. All relevant features of the input data, on the other hand, remain the same. As such, alterfactual explanations form the complement to counterfactual explanations - providing both explanation types should enable the user to grasp both relevant *and* irrelevant features. Mertes *et al.* already showed the potential in the concept of raising user awareness about irrelevant features [Mertes *et al.*, 2022b]. Nevertheless, due to the absence of an implementable solution, the researchers could only delve into the concept through the utilization of a fictional AI.

As such, in this work we introduce a GAN-based generation algorithm that is able to create both alterfactual and counterfactual explanations for image classifiers.¹ As alterfactual explanations convey completely different information than common methods, we investigate whether the understanding that users have of the explained AI system is also formed in a different way, or can even be improved. Our results show that alterfactual explanations outperform counterfactual explanations with regards to local model understanding.

2 Related Work

As the approach presented in this paper can be counted to the class of XAI methods that work by inducing counterfactual thinking processes, it is important to gain an understanding of how common methods from this field work. Therefore, this section gives an overview on related explanation concepts. Figure 1A illustrates the difference between those concepts using exemplary explanations for a fictional AI that decides if a person is creditworthy or not. We will use that scenario as a running example of how the different explanation paradigms would answer the question of *Why does the AI say that I am not creditworthy?*.

¹Our full implementation is open-source and available at <https://github.com/hcmlab/Alterfactuals>.

Factual Explanations - *There was another female person that also had rather little money, and she also did not get the credit.* - Factual explanations are the traditional way of explaining by example, and often provide a similar instance from the underlying data set (adapted or not) for the input data point that is to be explained [Keane *et al.*, 2021b]. Other approaches do not choose an instance from the dataset, but generate new ones [Guidotti *et al.*, 2019]. The idea behind factual explanations is that similar data instances lead to similar decisions, and the awareness of those similarities leads to a better understanding of the model. Further explanation mechanism that fall in this category are *Prototypical Explanations* and *Near Hits* Kim *et al.* [2016]; Herchenbach *et al.* [2022].

Counterfactual Explanations - *If you had that amount of money, you would get the credit.* - Counterfactual explanations are a common method humans naturally use when attempting to explain something and answer the question of *Why not ...?* [Miller, 2019; Byrne, 2019]. In XAI, they do this by showing a modified version of an input to an AI system that results in a different decision than the original input. Counterfactual explanations should be minimal, which means they should change as little as possible in the original input [Keane *et al.*, 2021b; Miller, 2021]. In certain scenarios, modern approaches for generating counterfactual explanations have shown significant advantages over feature attribution mechanisms (i.e., explanation approaches that highlight *which* features are important for a decision) in terms of mental model creation and explanation satisfaction [Mertes *et al.*, 2022a]. Wachter *et al.* [2017] name multiple advantages of counterfactual explanations, such as being able to detect biases in a model, providing insight without attempting to explain the complicated inner state of the model, and often being efficient to compute. For counterfactual explanations, a multitude of works exist that, similar to how we do it for alterfactual explanations, use GANs to automatically generate explanations for image classifiers Nemirovsky *et al.* [2022]; Van Looveren *et al.* [2021]; Khorram and Fuxin [2022]; Mertes *et al.* [2022a].

Semifactual Explanations - *Even if you had that amount of money, you would still not get the credit.* - Similar to counterfactual explanations, semifactual explanations are an explanation type humans commonly use. They follow the pattern of *Even if X, still P*, which means that even if the input was changed in a certain way, the prediction of the model would still not change to the foil [McCloy and Byrne, 2002]. In an XAI context, this means that an example, based on the original input, is provided that modifies the input in such a way that moves it toward the decision boundary of the model, but stops just before crossing it [Kenny and Keane, 2020]. Similar to counterfactual explanations, semifactual explanations can be used to guide a user’s future action, possibly in a way to deter them from moving toward the decision boundary [Keane *et al.*, 2021b].

3 Alterfactual Explanations

No matter what your gender is, the decision would not change. - The basic idea of alterfactual explanations investi-

gated in this paper is to strengthen the user’s understanding of an AI by showing irrelevant attributes of a predicted instance. Hereby, we define irrelevance as the property that the corresponding feature, regardless of its value, does not contribute in any way to the decision of the AI model. When looking at models that are making decisions by mapping some sort of input data $x \in X$ to output data $y \in Y$, the so-called *decision boundary* describes the region in X which contains data points where the corresponding y that is calculated by the model is ambiguous, i.e., lies just between different instances of Y . Thus, irrelevant features can be thought of as features that do not contribute to a data point’s distance to the decision boundary.

However, information that is carried out by explanations should be communicated as clearly as possible. Alterfactual explanations inform about the *irrelevance* of certain features - as such, it should be made clear that these features can take *any* possible value. If we would change the respective features only to a small amount, the irrelevance is not clearly demonstrated to the user. Therefore, an alterfactual explanation should change the affected features to the maximum amount possible. By doing so, they communicate that the feature, *even if it is changed as much as it can change*, still does not influence the decision. Thus, the definition of an alterfactual explanation is as follows:

Let $d : X \times X \rightarrow \mathbb{R}$ be a distance metric on the input space X . An *alterfactual explanation* for a model M is an altered version $a \in X$ of an original input data point $x \in X$, that maximizes the distance $d(x, a)$ whereas the distance to the decision boundary $B \subset X$ and the prediction of the model do not change: $d(x, B) = d(a, B)$ and $M(x) = M(a)$

Thus, the main difference between an alterfactual explanation and a counterfactual or semifactual explanation is that while the latter methods alter features resulting in a decreased distance to the decision boundary, our alterfactual method tries to keep that distance fixed. Further, while counterfactual and semifactual methods try to keep the overall change to the original input minimal [Keane *et al.*, 2021a; Kenny and Keane, 2020], alterfactual explanations do exactly the opposite, which is depicted in Figure 1B.

4 Generating Alterfactual Explanations

As we argue that alterfactual and counterfactual explanations convey different information, we designed a generative approach that is capable of creating both types of explanations in order to explain an image classifier. For both, a set of requirements arises that needs to be reflected in the objectives of our explanation generation approach.

1. The generated explanations should have high quality and look realistic.
2. The resulting explanation should be either classified as the same class as the original input (for alterfactual explanations), or as the opposite class (for counterfactual explanations).

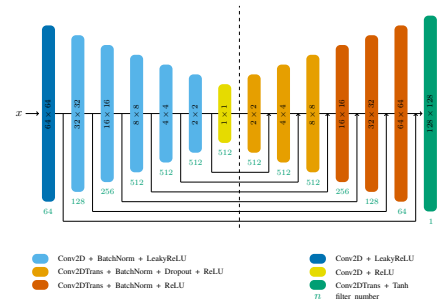


Figure 2: Architecture overview of the generator network.

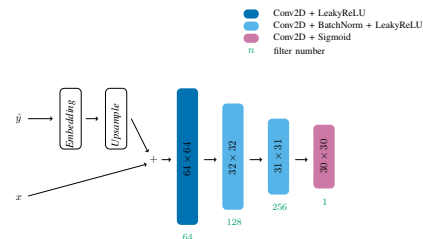


Figure 3: Architecture overview of the discriminator network.

3. For *alterfactual* explanations, the output image should change as much as possible, while for *counterfactual* explanations, it should change as little as possible.
4. For *alterfactual* explanations, only irrelevant features should change, i.e., the distance to the decision boundary should be maintained.

To address these objectives, different loss components (see next sections) were used to steer a GAN-based architecture to generate the desired explanations. A GAN-based approach was chosen as similar concepts have successfully been applied to the task of counterfactual explanation generation in various existing works [Olson *et al.*, 2021; Huber *et al.*, 2023; Nemirovsky *et al.*, 2022; Zhao, 2020; Mertes *et al.*, 2022a]. In order to allow for a more focused and comprehensive user study design, in this work, we focus on explaining a binary image classifier. However, although our specific generation architecture is designed for a binary classification problem, it would theoretically be possible to apply it to non-binary tasks by training separate models for each class vs. the union over all other classes. A schematic overview of our architecture can be seen in Figures 2 and 3. For a more detailed description, we refer to the appendix.

4.1 Adversarial Component

To address the first objective, an adversarial setting is used. Here, a generator network G is trained to take an original image x and a random noise vector z and transforms them into the respective explanation \hat{x} . As such, a mapping $\{x, z\} \rightarrow \hat{x}$ is learned by the generator. A discriminator network D is trained to identify the generated images as *fake* images in an adversarial manner.

Additionally, to partly target the second objective, we feed

a target class label $\hat{y} \in \{0, 1\}$ to the discriminator. By doing so, the discriminator learns not only to assess if the produced images are real or fake, but also has the capability to decide if an explanation fits the data distribution of the class it is supposed to belong to. A somewhat similar idea was put forth by Sharmanska *et al.* [2020] within the context of fairness and yielded promising results there. During training, the discriminator is alternately fed with real and fake data. For real data, the target class label \hat{y} reflects the class that the classifier to be explained assigns to the respective image x . For the generated explanations, the target class label \hat{y} reflects either the class that was assigned to the original image x (for alterfactual explanations), or the opposite class (for counterfactual explanations).

By letting the generator and discriminator compete against each other during training, it is enforced that the resulting images look realistic and resemble the data distribution of the respective target classes. The objective function for the adversarial setting is formulated as follows:

$$\mathcal{L}_{adversarial} = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x, \hat{y})] + \mathbb{E}_{x \sim p_{data}(x), z \sim p_{noise}(z)} [\log(1 - D(G(x, z), \hat{y}))] \quad (1)$$

4.2 Including Classifier Information

The second objective is further addressed by incorporating the decisions of the classifier to be explained into the generator’s loss function.

Let $C : X \rightarrow [0, 1]$ be a binary classifier with threshold 0.5. We define the classification target $\tilde{C}(x)$ as $\tilde{C}(x) := C(x)$ for alterfactual explanations and $\tilde{C}(x) := 1 - C(x)$ for counterfactual explanations. To measure the error between the actual classification of the generated explanation and the target classification, we used Binary Crossentropy (BCE) to define a classification loss \mathcal{L}_C :

$$\mathcal{L}_C = \mathbb{E}_{x, \hat{x} \sim p_{data}(x, \hat{x})} [\tilde{C}(x) \cdot \log C(\hat{x}) + (1 - \tilde{C}(x)) \cdot \log(1 - C(\hat{x}))] \quad (2)$$

4.3 SSIM Component

The third objective was addressed by including a similarity component into the loss function. Explanations are meant for humans. Therefore, using the Structural Similarity Index (SSIM) seemed to be an appropriate choice to measure image similarity for our approach, as it correlates with how humans are perceiving similarity in images [Wang *et al.*, 2004]. The parameters for SSIM were chosen as recommended by Wu *et al.* [2019].

As alterfactual explanations should change irrelevant features *as much as possible*, while counterfactual explanations should be *as close as possible* to the original image, the learning objective differs for both (low similarity for alterfactual explanations, high similarity for counterfactual explanations). With $[0, 1]$ as the range of SSIM, we designed the loss function as follows:

$$\mathcal{L}_{sim} = \begin{cases} \mathbb{E}_{x, \hat{x} \sim p_{data}(x, \hat{x})} [SSIM(x, \hat{x})] & \text{Alterfactual} \\ \mathbb{E}_{x, \hat{x} \sim p_{data}(x, \hat{x})} [1 - SSIM(x, \hat{x})] & \text{Counterfactual} \end{cases} \quad (3)$$

4.4 Feature Relevance Component

The fourth objective, i.e., forcing the network to only modify irrelevant features when generating alterfactual explanations, was addressed by using an auxiliary Support Vector Machine (SVM) classifier. Note that this loss is only applied when generating alterfactual explanations, not when generating counterfactual explanations. Li *et al.* [2018] and Elsayed *et al.* [2018] have shown theoretically and empirically that the last weight layer of a Neural Network converges to an SVM trained on the data transformed up to this layer if certain restrictions are met (e.g., the last two layers of the network have to be fully connected). An SVM’s decision boundary can be calculated directly - unlike the one of a Neural Network [Jiang *et al.*, 2018]. As such, we use an SVM which was trained to predict the classifier’s decision based on the activations of the classifier’s penultimate layer as a way to approximate the classifier’s decision boundary - if the generated alterfactual explanation has moved closer to the SVM’s separating hyperplane, relevant features were most likely modified. Although an unchanged decision boundary distance does not necessarily guarantee that no relevant features were modified, in our experiments, it was a good indicator.

The distance of x to the SVM’s separating hyperplane f was defined as follows, with w as the SVM’s weight vector:

$$SVM(x) = \left| \frac{f(x)}{\|w\|} \right| \quad (4)$$

The SVM loss is defined by the absolute difference in distance to the separating hyperplane between the original image and the generated alterfactual explanation:

$$\mathcal{L}_{SVM} = \mathbb{E}_{x \sim p_{data}(x), z \sim p_{noise}(z)} [|SVM(x) - SVM(\hat{x})|] \quad (5)$$

The final loss function is a summation of all the four loss components introduced above.

5 Evaluation Scenario

To assess the performance of our approach, we applied it to the Fashion-MNIST data set [Xiao *et al.*, 2017]. That data set contains 7,000 gray scale images for each of its ten categories of clothes, such as ‘ankle boots’ or ‘pullover’, split into *train* (6,000 images per class) and *test* (1,000 images per class) sets. The two classes we chose, ‘ankle boots’ and ‘sneakers’, were selected due to being somewhat similar in order not to oversimplify the classification task while still being distinct enough to be able to visually assess whether the generated explanations are clear. To create the classifier to be explained, we trained a relatively simple four-layer convolutional neural network, achieving an accuracy of 96.7% after 40 training epochs. The exact architecture and training configuration can be found in the appendix.

Our explanation generation architecture was trained for 14 epochs, until visually no further improvement could be observed. For alterfactual explanations, we reached a validity (i.e., which portion of the explanations are classified as the correct target class by the classifier) of 96.20% and an average SSIM of 0.32 (here, lower is better), whereas the counterfactual explanations reached a validity of 87.70% and an



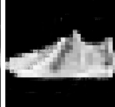








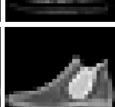
Ground Truth	AI Prediction	Original	Alterfactual	Counterfactual
Ankle Boot	Ankle Boot			
Sneaker	Sneaker			
Sneaker	Ankle Boot			
Ankle Boot	Sneaker			

Figure 4: Example outputs of our system. It can be seen that alterfactual explanations change features that are irrelevant to the classifier, e.g., the color of the shoes or the width of the boot shaft, while counterfactual explanations change relevant features like the presence or absence of a boot shaft. From top to bottom the original images are a correctly classified ankle boot and sneaker, followed by two inputs incorrectly classified as ankle boot and sneaker.

average SSIM of 0.90 (here, higher is better). For more details refer to the appendix. Exemplary generated explanations are shown in Figure 4. Note that, in order to verify if our alterfactual generation approach is applicable on a wider range of datasets, we additionally trained our approach on three other datasets: MNIST [LeCun, 1998], MaskedFace-Net [Cabani *et al.*, 2021], and a gray scale version of MaskedFace-Net. To further demonstrate that our approach can be adapted to be more model-agnostic and work without access to intermediate layers, we omitted the Feature Relevance component for those experiments. Training details, example outputs and computational results for those experiments can be found in the appendix.

6 User Study

6.1 Research Question and Hypotheses

We conducted a user study to validate whether the counterfactual and alterfactual explanations generated by our approach help human users to form correct model understanding of an AI system. Therefore, we only used results from the model trained on the Fashion-MNIST classifier in order to not overwhelm participants. To be able to compare our findings to existing work, we designed our study similar to Mertes *et al.* [2022b]. Our hypotheses are as follows:

- 1) Alterfactual and counterfactual explanations, as well as the combination of both, are more effective in enabling model understanding than no explanations.
- 2) There is a difference in model understanding and explanation satisfaction between alterfactual and counterfactual explanations, but we did not anticipate a specific direction since we see them as complementary concepts.
- 3) Compared to the individual explanations, a combination of alterfactual and counterfactual explanations is a more

effective way to enable a good model understanding and is more satisfying for users.

- 4) There is a difference between conditions regarding the understanding of relevant and irrelevant features, where alterfactual explanations are more effective to identify irrelevant features while counterfactual explanations should help more with identifying relevant features.

6.2 Methodology

Conditions and Explanation Presentation We used a between-groups design with four conditions. Participants in the *Control* condition were presented only with the original input images to the AI. No explanation was shown. In the *Alterfactual* and *Counterfactual* conditions, participants were presented with the original input images and either alterfactual or counterfactual explanations. In the *Combination* condition, participants were presented with the original input images as well as both the alterfactual and the counterfactual explanations.

Procedure The whole study was built using the *oTree* framework by Chen *et al.* [2016]. After answering questions about their demographic background, participants were given some general information about the data and their task during the prediction task. For the classifier, they were only told that an AI was trained to distinguish between ankle boots and sneakers. Two example images for each class (ankle boots and sneakers) were shown and some shoe specific terminology (e.g., "shaft") was introduced in order to make sure that participants have a common understanding of the terms they are asked about later on. Following this information, the participants were given an example input image for each class together with the classifier's prediction for this input image. In the explanation conditions, the participants were introduced to their corresponding explanation types (counterfactuals, alterfactuals or a combination) and could explore the explanations for those two images. After that, each participant answered a quiz about the information that was given up to that point, to make sure that they understood everything correctly. Subsequently, the study itself started. It was divided into three parts: For assessing the participants' understanding of the classifier, we used (i) a prediction task for assessing the local understanding, i.e., to assess if the participants understand why the AI makes a *specific* decision, and (ii) a questionnaire about the relevance of certain features for assessing the global understanding, i.e., to assess if the participants understand how the AI works *overall*. To assess the participants' explanation satisfaction, we used (iii) an explanation satisfaction questionnaire. The three phases of the experiment are described below.

Local Model Understanding: Prediction Task To measure the local understanding of the classifier, we used a prediction task, which assesses the participants' ability to anticipate the AI classifier's decisions [Hoffman *et al.*, 2018]. Eight examples were shown, covering all possible classification outcomes (two correctly classified images for both sneakers and ankle boots, and two incorrectly classified images for both) to avoid bias. Figure 4 shows four of the images from the study. The example images were chosen randomly but

we made sure that the alterfactual and counterfactual explanations generated by our model for those images were valid (i.e., the classifier predicted the same class as for the original image when fed with the alterfactual explanation, and the opposite class when fed with the counterfactual explanation). Participants had to predict the classifier’s decision for each example image. Participants were additionally asked about their own opinion on which class the original shoe image belonged to. The answers to that particular question were not further analyzed - it was only added to help the participants distinguish between their own opinion and their understanding of the classifier. After predicting an example, they were told the correct label and the AI classifier’s decision before moving on to the next example. The order of the examples was randomized.

Global Model Understanding: Feature Relevance While the Prediction Task can be seen as *local* measurement of the users’ understanding of the model in specific instances, we also wanted to investigate whether participants understood the *global* relevance of different features. To this end, we looked at two features that were relevant for our classifier (“presence/absence of a boot shaft” and “presence/absence of an elevated heel”) as well as two features that were irrelevant for our classifier (“boot shaft width” and “the shoe’s color and pattern on the surface area”). These features were chosen based on the authors’ experience from training the classifier and a-priori explorations with the Feature Attribution explanation mechanisms LIME [Ribeiro *et al.*, 2016] and SHAP [Lundberg and Lee, 2017]. Note that, although the classifier is still a black box and there is no definitive proof that the chosen features reflect the classifier’s inner workings entirely accurately, we decided that using those mechanisms for the feature choice are the best proxy that we have. As such, after the participants went through the eight examples that were used for the prediction task, they were asked for each feature how much they agreed that it was relevant to the AI’s decisions on a 5-point Likert scale (0 = strongly disagree, 4 = strongly agree).

Explanation Satisfaction In order to measure the participants’ subjective satisfaction, we used the Explanation Satisfaction Scale proposed by Hoffman *et al.* [2018] which consists of eight items rated on a 5-point Likert scale (0 = strongly disagree, 5 = strongly agree) that we averaged over all items. Since it does not apply to our use-case, we excluded the 5th question of the questionnaire. The seven remaining items address *confidence*, *predictability*, *reliability*, *safety*, *wariness*, *performance*, *likeability*. Finally, the participants had the possibility to give free text feedback.

6.3 Participants

Through a power analysis, we estimated a required sample size of at least 21 per condition for a MANOVA with 80% power and an alpha of 5%, based on the Pillai’s Trace of 0.13 reported for the study by Mertes *et al.* [2022b]. 131 Participants between 18 and 29 years ($M = 22.2$, $SD = 2.44$) were recruited at the University of *blinded for review*. 61 of them were male, 70 female. The participants were randomly separated into the four conditions (33 per condition and 32 in the

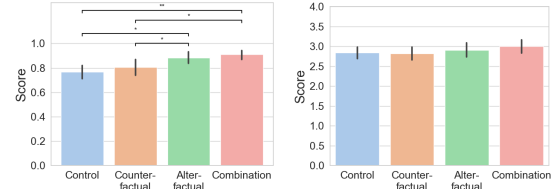


Figure 5: Left: Mean participant prediction accuracy of the AI’s prediction by condition. The conditions containing alterfactual explanations outperformed all other conditions. Right: Mean understanding of the irrelevant and relevant features in our study. Error bars represent the 95% CI. * $p < .05$, ** $p < .001$.

Alterfactual condition). The highest level of education that most participants held (76.3%) was a high-school diploma. Only 11.5% of the participants had no experience with AI. Most of the participants (74%) have heard from AI in the media. Excluding participants that had no opinion on the subject, the participants expected a positive impact of AI systems in the future ($M = 3.73$ on a 5-point Likert Scale from 1 = “Extremely negative” to 5 = “Extremely positive”). There were no substantial differences in the demographics between conditions (see appendix).

7 Results

7.1 Model Understanding

To investigate the impact of the four different experimental conditions on the (1) feature understanding and (2) prediction accuracy, we conducted a MANOVA. We found a significant difference, Wilks’ Lambda = 0.859, $F(6,252) = 3.31$, $p = .004$.

The following ANOVA revealed that **only the prediction accuracy of the participants showed significant differences between the conditions**:

- *Feature Understanding*: $F(3,127) = 0.877$, $p = .455$.
- *Prediction Accuracy*: $F(3,127) = 6.578$, $p < .001$.

As displayed in Figure 5, the post-hoc t-tests showed that the participants’ prediction accuracy was significantly better in the *Alterfactual* and *Combination* conditions compared to the other conditions. The effect size d is calculated according to Cohen [2013]:

- *Alterfactual vs. Control*: $t(127) = 3.19$, $p = .002$, $d = 0.79$ (medium effect).
- *Alterfactual vs. Counterfactual*: $t(127) = 2.06$, $p = .042$, $d = 0.51$ (medium effect).
- *Combination vs. Control*: $t(127) = 3.93$, $p < .001$, $d = 0.97$ (large effect).
- *Combination vs. Counterfactual*: $t(127) = 2.79$, $p = .006$, $d = 0.69$ (medium effect).

These results regarding the prediction task confirm our hypothesis that the **conditions with alterfactual explanations outperform the condition without explanations in the prediction task**. Further, **the combination of both explanation types did significantly outperform counterfactual explanations**. However, our hypothesis that the combination is

more effective in terms of enabling a correct model understanding than alterfactual explanations has to be rejected.

7.2 Relevant and Irrelevant Information

As reported in the section above, we did not find a significant overall difference in the feature understanding task (see Figure 5). However, in order to investigate our hypotheses about irrelevant vs. relevant features, we conducted another MANOVA between the conditions and the combined understanding values for the two relevant features and the two irrelevant features. This MANOVA did not find any significant differences, Wilks' Lambda = 0.951, $F(6,252) = 1.07$, $p = .379$. The mean understanding per condition can be found in the appendix.

7.3 Explanation Satisfaction

The ANOVA revealed that there were no significant differences in the subjective explanation satisfaction between the three explanation conditions, $F(2,95) = 0.34$, $p = .713$. The mean satisfaction values with standard deviation were: *Counterfactual* condition: 3.54 ± 0.53 ; *Alterfactual* condition: 3.65 ± 0.6 ; *Combination* condition: 3.58 ± 0.5 .

8 Discussion

With our proposed GAN-based approach, we demonstrated that it is possible to generate both counterfactual and alterfactual explanations for a black box image classifier. Using computational metrics, we showed that both of those generated explanations fulfill their respective requirements: The counterfactual explanations are very similar to the original images (i.e., 0.90 average SSIM) but change the classifiers prediction in 87.70% of the cases while alterfactual explanations are very different from the original image (i.e., 0.32 average SSIM), but do not change the classifier's prediction in 96.20% of the cases. For the prediction task of our user study, alterfactual explanations and the combination of alterfactual and counterfactual explanations performed significantly better than the other two conditions demonstrating the potential of alterfactual explanations to facilitate local model understanding. However, we did not observe a significant difference for the feature relevance understanding. This is highly interesting, as it contrasts with a previous study by Mertes *et al.* [2022b]. There, a similar experimental design was employed for assessing the effect that alterfactual explanations have on users' mental models of a hard-coded classifier that assesses numerical feature descriptors for a fictional classification problem. In contrast to our work, they neither used a real classifier nor an alterfactual generation algorithm, but only mock-up decisions and explanations. In their scenario, alterfactual explanations led to a significantly better feature relevance understanding, while not having a substantial impact on the performance in a prediction task. A possible explanation for this is the fact that our study was conducted in the context of fashion classification, where the users might already have had a quite distinctive mental model of the problem domain itself. Further, images might be more accessible than numerical feature descriptors to end users. As such, the global understanding of the classifier might already be positively biased. This argument is supported by looking at the

feature relevance understanding results of the control group - although not seeing any explanations, they already performed very well in identifying relevant features. However, as can be seen by the significant performance improvement in the prediction task, the local understanding of the model does not necessarily benefit from the identification of globally relevant features. As the classification model is imperfect, a global understanding of the use case itself does not necessarily imply an understanding of cases, e.g. when the classifier's decision does not correctly model reality. Furthermore, our results regarding the global model understanding should be taken with a grain of salt, since the fashion-classifier is a black-box model - even though we used post-hoc explanation methods (SHAP and LIME), we cannot be certain that our choice of features is completely accurate. Interestingly, we did not observe any significant differences in explanation satisfaction. This indicates that participants felt similarly satisfied by all explanation methods even though the alterfactual and combined explanations objectively helped more during the prediction task. The presentation of more information (i.e., in the combination condition) could have led to a higher cognitive load and influenced the subjective assessments of explanation satisfaction, resulting in the difference between objective measurement (i.e., model understanding) and subjective measurement (i.e., explanation satisfaction).

9 Conclusion

In this paper, we demonstrate the practical feasibility of a recently proposed concept for explaining AI models called *alterfactual explanations* that *alter* as much irrelevant information as possible while maintaining the distance to the decision boundary.

We show for the first time that it is possible to generate such explanations for black box models and briefly evaluated them computationally. Furthermore, we showed in a user study that our generated alterfactual explanations can complement counterfactual explanations. In that study, we compared how users' model understanding of a binary image classifier changes when being confronted with counterfactual explanations, alterfactual explanations, or a combination of both. Further, a control group was assessed that did not see any explanations. We found that in a prediction task, where the classifier's prediction had to be anticipated by looking at the explanations, users performed significantly better when they were provided with explanations that included alterfactual explanations compared to users that did not see alterfactual explanations, although we did not observe a significant difference in explanation satisfaction.

Overall, we showed that alterfactual explanations are a promising explanation method that can complement counterfactual explanations in future XAI systems.

10 Acknowledgments

This research was partially funded by the DFG through the Leibniz award of Elisabeth André (AN 559/10-1).

References

- Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bannetot, Siham Tabik, Alberto Barabado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.
- Ruth MJ Byrne. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *IJ-CAI*, pages 6276–6282, 2019.
- Adnane Cabani, Karim Hammoudi, Halim Benhabiles, and Mahmoud Melkemi. Maskedface-net—a dataset of correctly/incorrectly masked face images in the context of covid-19. *Smart Health*, 19:100144, 2021.
- Daniel L Chen, Martin Schonger, and Chris Wickens. otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97, 2016.
- Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.
- Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. *Advances in neural information processing systems*, 31, 2018.
- Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6):14–23, 2019.
- Marvin Herchenbach, Dennis Müller, Stephan Scheele, and Ute Schmid. Explaining image classifications with near misses, near hits and prototypes: Supporting domain experts in understanding decision boundaries. In *International Conference on Pattern Recognition and Artificial Intelligence*, pages 419–430. Springer, 2022.
- Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.
- Tobias Huber, Maximilian Demmler, Silvan Mertes, Matthew L. Olson, and Elisabeth André. Ganterfactual-rl: Understanding reinforcement learning agents’ strategies through visual counterfactual explanations. In Noa Agmon, Bo An, Alessandro Ricci, and William Yeoh, editors, *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2023, London, United Kingdom, 29 May 2023 - 2 June 2023*, pages 1097–1106. ACM, 2023.
- Yiding Jiang, Dilip Krishnan, Hossein Mobahi, and Samy Bengio. Predicting the generalization gap in deep networks with margin distributions. *arXiv preprint arXiv:1810.00113*, 2018.
- Mark T Keane, Eoin M Kenny, Eoin Delaney, and Barry Smyth. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. *arXiv preprint arXiv:2103.01035*, 2021.
- Mark T Keane, Eoin M Kenny, Mohammed Temraz, Derek Greene, and Barry Smyth. Twin systems for deepcbr: A menagerie of deep learning and case-based reasoning pairings for explanation and data augmentation. *arXiv preprint arXiv:2104.14461*, 2021.
- Eoin M Kenny and Mark T Keane. On generating plausible counterfactual and semi-factual explanations for deep learning. *arXiv preprint arXiv:2009.06399*, 2020.
- Saeed Khorram and Li Fuxin. Cycle-consistent counterfactuals by latent transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10203–10212, 2022.
- Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.
- Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- Yu Li, Lizhong Ding, and Xin Gao. On the decision boundary of deep neural networks. *arXiv preprint arXiv:1808.05385*, 2018.
- Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- Rachel McCloy and Ruth MJ Byrne. Semifactual “even if” thinking. *Thinking & Reasoning*, 8(1):41–67, 2002.
- Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth André. Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. *Frontiers in artificial intelligence*, 5, 2022.
- Silvan Mertes, Christina Karle, Tobias Huber, Katharina Weitz, Ruben Schlagowski, and Elisabeth André. Alterfactual explanations—the relevance of irrelevance for explaining ai systems. *arXiv preprint arXiv:2207.09374*, 2022.
- Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.
- Tim Miller. Contrastive explanation: A structural-model approach. *The Knowledge Engineering Review*, 36, 2021.
- Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu, and Abhishek Gupta. Countergan: Generating counterfactuals for real-time recourse and interpretability using residual gans. In *Uncertainty in Artificial Intelligence*, pages 1488–1497. PMLR, 2022.

Matthew L. Olson, Roli Khanna, Lawrence Neal, Fuxin Li, and Weng-Keen Wong. Counterfactual state explanations for reinforcement learning agents via generative deep learning. *Artif. Intell.*, 295:103455, 2021.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

Stuart Russell and Peter Norvig. Artificial intelligence: A modern approach global edition. *Pearson*, 2016.

Viktoriia Sharmanska, Lisa Anne Hendricks, Trevor Darrell, and Novi Quadrianto. Contrastive examples for addressing the tyranny of the majority. *arXiv preprint arXiv:2004.06524*, 2020.

Peter Stone, Rodney Brooks, Erik Brynjolfsson, Ryan Calo, Oren Etzioni, Greg Hager, Julia Hirschberg, Shivaram Kalyanakrishnan, Ece Kamar, Sarit Kraus, et al. One hundred year study on artificial intelligence: Report of the 2015-2016 study panel. Technical report, Technical report, Stanford University, 2016.

Arnaud Van Looveren, Janis Klaise, Giovanni Vacanti, and Oliver Cobb. Conditional generative models for counterfactual explanations. *arXiv preprint arXiv:2101.10123*, 2021.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Yifan Wu, Fan Yang, Yong Xu, and Haibin Ling. Privacy-protective-gan for privacy preserving face de-identification. *Journal of Computer Science and Technology*, 34:47–60, 2019.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Baobao Zhang and Allan Dafoe. Artificial intelligence: American attitudes and trends. *Available at SSRN 3312874*, 2019.

Yunxia Zhao. Fast real-time counterfactual explanations. *arXiv preprint arXiv:2007.05684*, 2020.

A GAN Architecture and Training

Generator Model

The GAN’s generator architecture is listed in Table 1.

Discriminator Model

The GAN’s discriminator architecture is listed in Table 2.

A.1 Training Configuration and Hyperparameters

The training configuration and hyperparameters are shown in Table 3. The Adam optimizer was configured with $\beta_1 = 0.5$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$.

Further, the Support Vector Machine (SVM) that was included in the loss function (see main paper) was trained with the parameters listed in Table 4.

B Classifier Architecture and Training

In Table 5, the model architecture for the classifier that we used in our evaluation scenario is described. The training configuration and hyperparameters are shown in Table 6. The Adam optimizer was configured with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e - 8$.

C Additional Dataset Experiments

In order to demonstrate that our alterfactual generation approach is generalizable to different datasets, we additionally trained models for three other datasets. Here, we omitted the Feature Relevance component. As for that component an additional SVM has to be trained on the penultimate layer of the classifier layer, it takes away the model-agnostic property from the alterfactual generation network. By performing these additional experiments, we show that the approach can simply be adapted to be model-agnostic, although that may negatively affect the outcomes of the results - it is not specifically forced that *only* irrelevant features change. For the classifiers, we used the same architecture as for the Fashion-MNIST dataset, although batch size and epochs were modified to fit the hardware that we used.

C.1 MNIST

As the MNIST datasets has more than two classes (each class contains hand-drawn images of one specific digit), we picked the two digits that are most likely to be confused by deep learning classifiers: *Three* and *Eight*. The MNIST classifier was trained for 9 epochs with batch size 32. Besides not using the Feature Relevance component and increasing the epoch number to 42, the GAN network was trained with the same parameter settings as for the Fashion-MNIST dataset. We reached a validity of 95.92% and an average SSIM of 0.425. Example outputs are shown in Figure 6.

C.2 MaskedFace-Net

The MaskedFace-Net dataset contains images of people wearing face masks. Binary labels are provided, indicating that on the respective image the mask is worn correctly or incorrectly. The classifier was trained for 2 epochs with batch size 128. Besides not using the Feature Relevance component and decreasing the epoch number to 11, the GAN network was trained with the same parameter settings as for the Fashion-MNIST dataset. We reached a validity of 84.27% and an average SSIM of 0.091. Example outputs are shown in Figure 7.

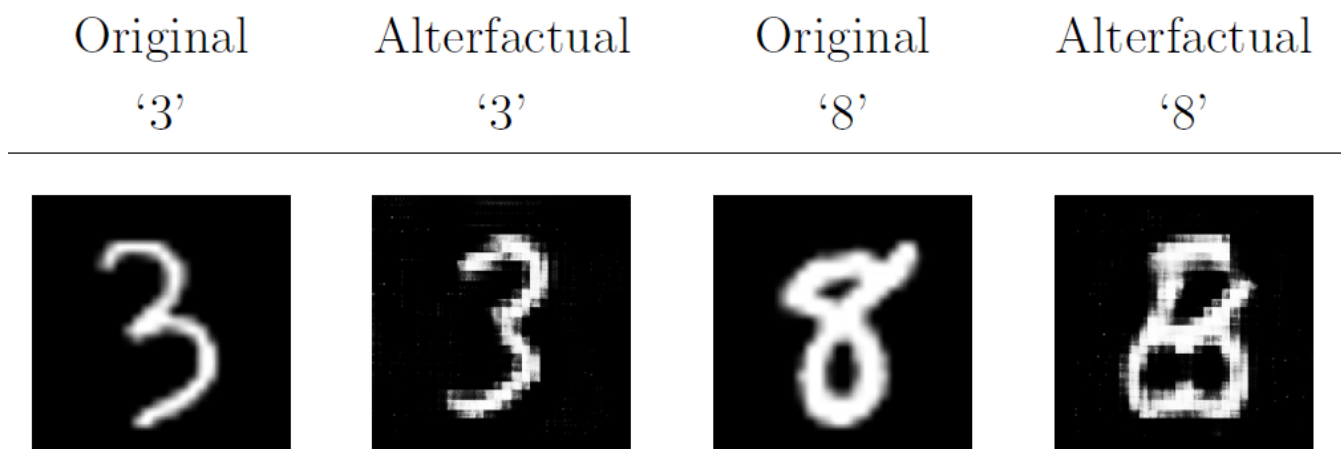


Figure 6: Exemplary alterfactual outputs for the MNIST dataset.

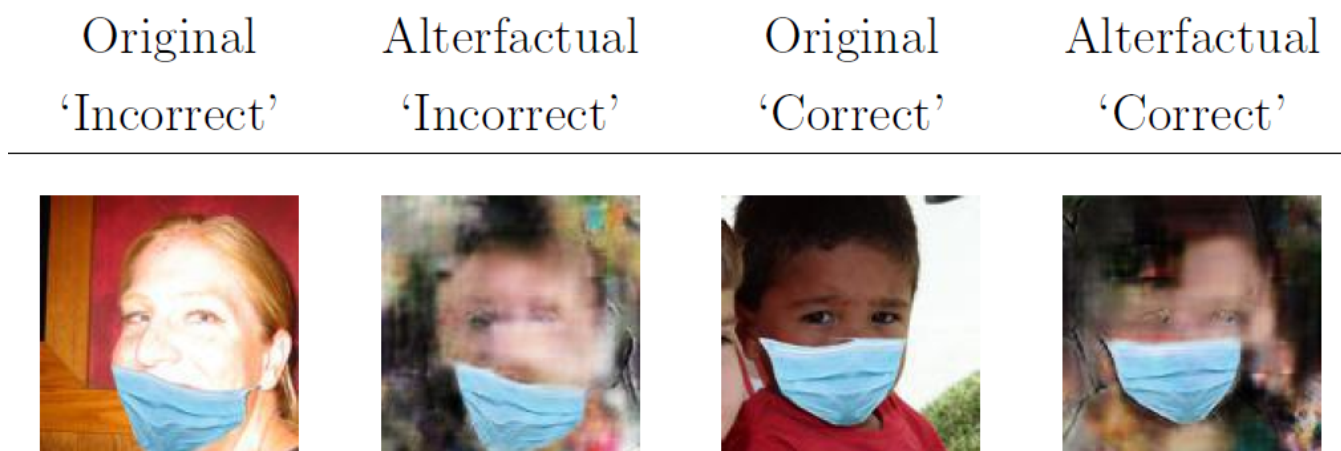


Figure 7: Exemplary alterfactual outputs for the MaskedFace-Net dataset.

C.3 MaskedFace-Net (Gray Scale)

Here, we also used the MaskedFace-Net dataset, but converted it to gray scale, demonstrating that our approach also works with gray scale data. The classifier was trained for 1 epoch with batch size 128. Besides not using the Feature Relevance component and decreasing the epoch number to 6, the GAN network was trained with the same parameter settings as for the Fashion-MNIST dataset. We reached a validity of 48.89% and an average SSIM of 0.002. Example outputs are shown in Figure 8.

D User Study

D.1 Demographic Details

The mean age and education level, as well as the percentage of female participants, per condition can be seen in Table 7. For the AI experience and Attitude we adapted a description of AI from Zhang and Dafoe [2019] and Russell and Norvig [2016] to “The following questions ask about Artificial Intelligence (AI). Colloquially, the term ‘artificial intelligence’ is often used to describe machines (or computers) that mimic ‘cognitive’ functions that humans associate with the human mind, such as ‘learning’ and ‘problem solving’.” After this description, participants had to select one or more item describing their experience with AI. The distribution of the items for each condition is shown in Fig. 9. Following this we adapted a question from Zhang and Dafoe [2019] to measure the participants’ attitude towards AI. We asked them to rate their answer to the question “Suppose that AI agents would achieve high-level performance in more areas one day. How positive or negative do you expect the overall impact of such AI agents to be on humanity in the long run?” on a 5-point Likert scale from “Extremely negative” to “Extremely positive”. The participants also had the option to answer “I do not know” here, which would exclude them from the evaluation of this question. The mean results for each condition are shown in Table 7.

E Additional Post-Hoc Results

For completeness, we also report the results of the post-hoc t-tests on the participants’ prediction accuracy that were not significant. The effect size d is calculated according to Cohen [2013]:

- **Counterfactual vs. Control:** $t(127) = 1.14$, $p = .258$, $d = 0.28$
- **Combination vs. Alterfactual:** $t(127) = 0.71$, $p < .478$, $d = 0.18$.

For feature understanding and explanation satisfaction we did not calculate post-hoc tests since the ANOVA was not significant.

E.1 Mean Understanding of Features

Figure 10 shows the mean understanding (as assessed by the feature understanding task) per condition.

E.2 Explanation Satisfaction Scale

For evaluating explanation satisfaction, we used the Explanation Satisfaction scale by Hoffmann [Hoffman *et al.*, 2018] except one item that did not apply to our use case. The items that we used were as follows, where each item was rated on a 5-point likert scale (1 = strongly disagree, 5 = strongly agree):

- From the explanations, I **understand** how the AI makes its decision.
- The explanations of how the AI makes its decision are **satisfying**.
- The explanations of how the AI makes its decision have **sufficient detail**.
- The explanations of how the AI makes its decision seem **complete**.
- The explanations of how the AI makes its decision are **useful** to predict the AI’s decision.
- The explanations of how the AI makes its decision show me how **accurate** the AI is.
- The explanations let me judge when I should **trust and not trust** the AI.

E.3 Study Design

Screenshots of the user study are shown in Figures 11 to 40. The *Combination* condition is shown.

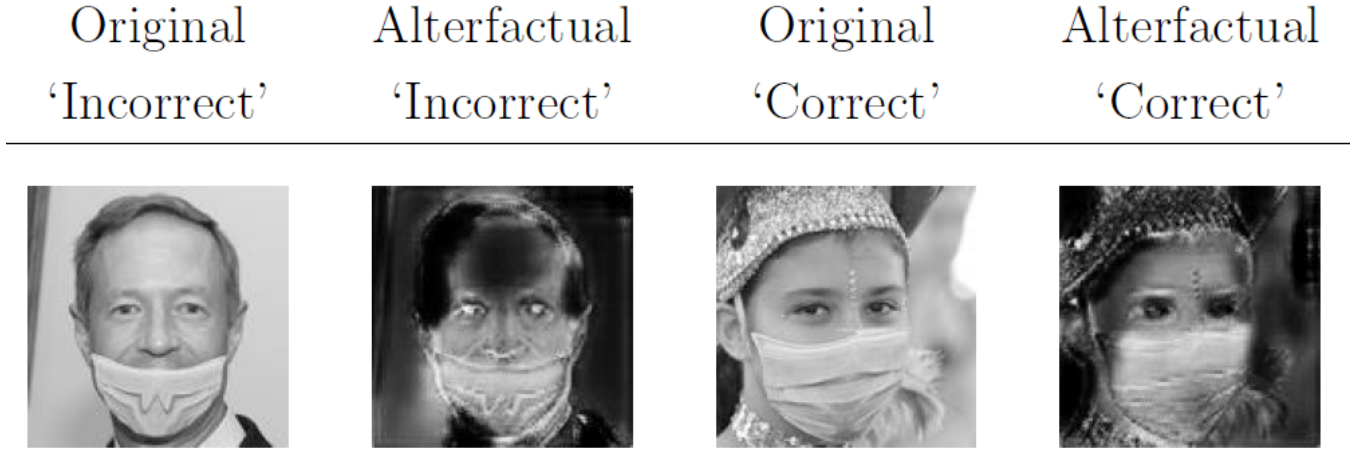


Figure 8: Exemplary alterfactual outputs for the gray scale version of the MaskedFace-Net dataset. It can be clearly seen that the only part of the image that gets unchanged is the mask itself - indicating that everything else is irrelevant.

Layer	Description	# Filter	Size	Stride	Dropout	BatchNorm	Activation
1	Conv2D	64	4x4	2	-	no	LeakyReLU (0.2)
2	Conv2D	128	4x4	2	-	yes	LeakyReLU (0.2)
3	Conv2D	256	4x4	2	-	yes	LeakyReLU (0.2)
4	Conv2D	512	4x4	2	-	yes	LeakyReLU (0.2)
5	Conv2D	512	4x4	2	-	yes	LeakyReLU (0.2)
6	Conv2D	512	4x4	2	-	yes	LeakyReLU (0.2)
7	Conv2D	512	4x4	2	-	no	ReLU
8	Conv2DTranspose	512	4x4	2	0.5	yes	ReLU
9	Conv2DTranspose	512	4x4	2	0.5	yes	ReLU
10	Conv2DTranspose	512	4x4	2	0.5	yes	ReLU
11	Conv2DTranspose	256	4x4	2	-	yes	ReLU
12	Conv2DTranspose	128	4x4	2	-	yes	ReLU
13	Conv2DTranspose	64	4x4	2	-	yes	ReLU
14	Conv2DTranspose	1	4x4	2	-	no	Tanh

Table 1: Generator Architecture used in our evaluation scenario. The architecture is adapted from Wu *et al.* [2019]. Where BatchNorm, Dropout, or Activation function occurred together, the order applied was BatchNorm - Dropout - Activation.

Layer	Description	# Filter	Size	Stride	BatchNorm	Activation
0a	Embedding	-	8x8	-	no	-
0b	Upsample	-	128x128	-	no	-
1	Conv2D	64	4x4	2	no	LeakyReLU (0.2)
2	Conv2D	128	4x4	2	yes	LeakyReLU (0.2)
3	Conv2D	256	4x4	2	yes	LeakyReLU (0.2)
4	Conv2D	1	4x4	2	no	Sigmoid

Table 2: Discriminator Architecture used in our evaluation scenario. Where BatchNorm and Activation function occurred together, Batch-Norm preceded the activation function. The first two layers, marked as '0a' and '0b' were used to upsample the label information to the size of the input image. The label and image were passed together to layer 1. The architecture is adapted from Wu *et al.* [2019].

Batch Size	1
Epochs	14
Learning Rate Generator	1e-4
Learning Rate Discriminator	1e-4
Optimizer	Adam

Table 3: The setting used to train the GAN.

C (Regularisation)	10
Kernel	linear
Iterations	5000

Table 4: The setting used to train the SVM.

Layer	Description	# Filter	Size	Stride	BatchNorm	Activation
1	Conv2D	32	3x3	1	yes	ReLU
2	Conv2D	32	3x3	1	yes	ReLU
3	MaxPool2D	-	2x2	2	no	-
4	Conv2D	64	3x3	1	yes	ReLU
5	Conv2D	64	3x3	1	yes	ReLU
6	GAP	-	-	-	no	-
7	Dense	-	2	-	no	Softmax

Table 5: Classifier architecture used to train the classifier for the MNIST-Fashion dataset (classes *Sneaker* and *Ankle Boot*). Where BatchNorm and Activation function occurred together, BatchNorm preceded the activation function.

Batch Size	32
Epochs	40
Learning Rate	1e-3
Optimizer	Adam
Loss Function	Binary Cross Entropy

Table 6: The setting used to train the Fashion-MNIST classifier.

	Control	Counterfactual	Alterfactual	Combination
Mean age	22.0	22.5	21.6	22.6
Percentage of female participants	52	58	47	58
Highest level of Education	2.27	2.42	2.19	2.21
Mean AI Attitude	3.50	3.82	3.84	3.75

Table 7: Demographic data across conditions. The highest level of education was measured as follows: 1 - No education, 2 - High school graduation, 3 - Vocational training, 3 - Bachelor, 4 - Master, 5 - Doctor. The Attitude towards AI is measured on 5-Point Likert scale from “Extremely negative” to “Extremely positive”.

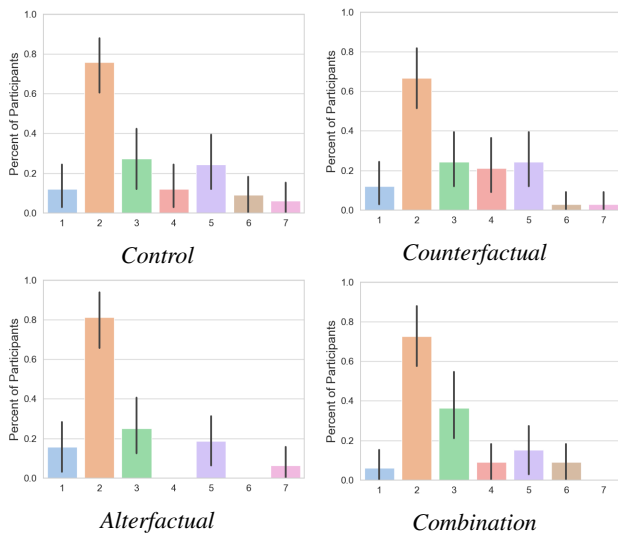


Figure 9: Distribution of the chosen AI experience items for each condition. The x-axis depicts the following items: 1 - I do not have any experience in AI related topics; 2 - I know AI from the media; 3 - I use AI technology in my private life; 4 - I use AI technology in my work; 5 - I have taken at least one AI related course; 6 - I do research on AI-related topics; 7 - Other:

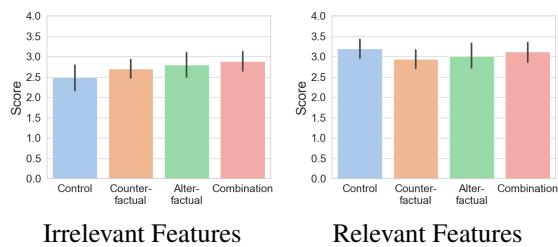


Figure 10: Mean understanding of the irrelevant and relevant features in our study. Error bars represent the 95% CI.

3%

Personal Information

What is your age?

Gender: How do you identify?

☐ Man

☐ Woman

☐ Other

Which is the highest level of education you have ?

☐ No education

☐ School graduation

☐ Completed vocational training

☐ Bachelor

☐ Master

☐ Doctor

Next

Figure 11: Screenshot of the user study, part 1.

7%

Personal Information

The following questions are about Artificial Intelligence (AI). Colloquially, the term "artificial intelligence" is often used to describe machines (or computers) that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving".

Do you have experience with AI? Check all that apply.

☐ I do not have any experience in AI related topics.

☐ I know AI from the media.

☐ I use AI technology in my private life.

☐ I use AI technology in my work.

☐ I have taken at least one AI related course.

☐ I do research on AI related topics.

☐ Other

Next

Figure 12: Screenshot of the user study, part 2.

10%

Personal Information

The following questions are about Artificial Intelligence (AI). Colloquially, the term "artificial intelligence" is often used to describe machines (or computers) that mimic "cognitive" functions that humans associate with the human mind, such as "learning" and "problem solving".

AI agents are already able to perform some complex tasks. Examples for such intelligent agents are search engines, chatbots, chessbots, and voice assistants. Suppose that AI agents would achieve high-level performance in more areas one day.

	Extremely negative	Negative	Neutral	Positive	Extremely positive	I don't know
How positive or negative do you expect the overall impact of such AI agents to be on humanity in the long run?	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Next

Figure 13: Screenshot of the user study, part 3.

14%

Start

You are now entering the actual study. It consists of the following parts:

1. Introduction to the domain & survey mechanics
2. Quiz on domain & survey mechanics
3. Predict the AI's prediction aided by explanations
4. Quiz on how the AI decides
5. Questions regarding your satisfaction with the explanations

Next

Figure 14: Screenshot of the user study, part 4.

17%

Ankle Boots & Sneakers

Read the following information carefully. You will be quizzed on it later.

In this study, you will be shown some images of shoes. Some of the images you will see show **ankle boots**, the others will show **sneakers**. Ankle boots are shoes that cover the whole foot and reach up to the ankle. Here, they may also cover parts of the leg.

We have trained an AI to predict whether the shoes that can be seen in the images are ankle boots or sneakers. Later in this survey, you will have to predict the AI's decisions. Specifically, we will show you some images that are classified by the AI and some additional information to help you understand **why** the AI made its decision, but we will not tell you upfront **which** decision it made.

In total you will have to make **8 predictions** in this study.

The following section will give you a brief overview of what these images look like.

The pictures you will see show either an ankle boot or a sneaker.

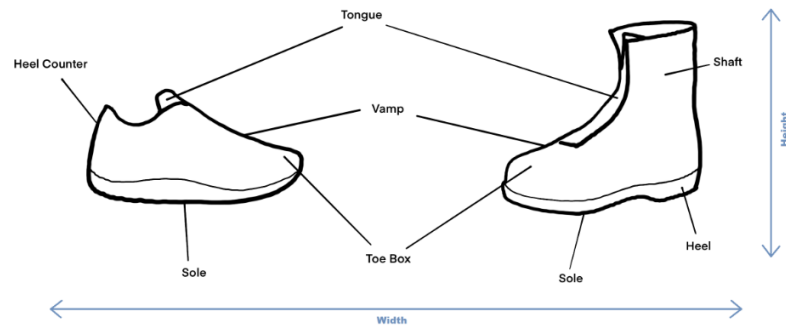
Examples of **ankle boots** are shown in the following pictures:



Examples of **sneakers** are shown in the following pictures:



You will be asked about specific parts of shoes/boots later in this study. The following terminology will be used (it will be shown again later):



Next

Figure 15: Screenshot of the user study, part 5.

21%

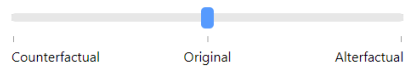
Ankle Boots & Sneakers I

During this study, you will be shown various images of shoes. We have trained an AI that decides whether these images show ankle boots or sneakers. **The AI is not perfect and can make mistakes.**

To aid you in your analysis of the AI, we will provide you with two additional explanations of the AI's prediction. Below you find an image with a slider. By moving the slider to the **right**, the image will change to a so-called **alterfactual** image. This alterfactual image shows you how the original image could be **maximally** modified **without influencing** the AI's prediction. This means that ideally only **irrelevant** attributes (for this particular decision) will be altered.

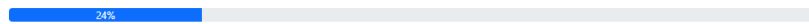
By moving the slider to the **left** the image will change to a so-called **counterfactual** image. This counterfactual image shows you how the original image could be **minimally** modified to **change** the AI's prediction. This means that ideally only **relevant** attributes (for this particular decision) will be altered

In **this** example, the AI classified the image as **ankle boot**. Thus, by moving the slider to the right, the image will change to show you a modified version of the original image which the AI would still classify as ankle boot, even though it has been maximally altered (e.g. alterfactual image). By moving the slider to the left the image change to show you a modified version of the original image which the AI would classify as sneaker (e.g. counterfactual image). Try it out!



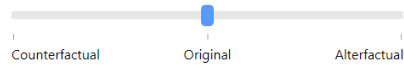
Next

Figure 16: Screenshot of the user study, part 6.



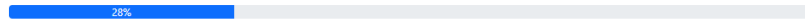
Ankle Boots & Sneakers II

Below you find an image with a slider. In **this** example, the AI classified the image as **sneaker**. Thus, by moving the slider to the **right**, the image will change to show you a modified version of the original image which the AI would still classify as sneaker, even though it has been maximally altered (e.g. alterfactual image). By moving the slider to the **left** the image will change to show you a modified version of the original image which the AI would classify as ankle boot (e.g. counterfactual image). Try it out!



Next

Figure 17: Screenshot of the user study, part 7.



Quiz

Please answer the following questions according to the information you have been shown in the last few pages.

Be careful: Answering incorrectly can exclude you from the remaining survey! Before pressing the "Next"-Button, ask your supervisor to give feedback.

The shoes are always shown from the following position:

- ☐ from above
- ☐ from below
- ☐ from the front
- ☐ from the side
- ☐ from the back

Which of the following statements is true?

- ☐ The AI is always right.
- ☐ The AI is not perfect. It can be wrong sometimes.

When the slider is in the **middle**, what does the image show you?

- ☐ They show the original image that was classified by the AI
- ☐ They show how the original image could maximally be modified without influencing the AI's prediction
- ☐ They show how the original image could be minimally modified to change the AI's prediction
- ☐ They always show a image classified as a sneaker
- ☐ They always show a image classified as an ankle boot

When the slider is on the **left**, what does the image show you?

- ☐ They show the original image that was classified by the AI
- ☐ They show how the original image could maximally be modified without influencing the AI's prediction
- ☐ They show how the original image could be minimally modified to change the AI's prediction
- ☐ They always show a image classified as a sneaker
- ☐ They always show a image classified as an ankle boot

When the slider is on the **right**, what does the image show you?

- ☐ They show the original image that was classified by the AI
- ☐ They show how the original image could maximally be modified without influencing the AI's prediction
- ☐ They show how the original image could be minimally modified to change the AI's prediction
- ☐ They always show a image classified as a sneaker
- ☐ They always show a image classified as an ankle boot

Next

Figure 18: Screenshot of the user study, part 8.

Prediction 1/8

Below you see an image that was classified by the AI. To aid you in your analysis of the AI, we will provide you with an additional explanation. If you move the slider below the image to the right, you will see how the image could be maximally modified **without influencing** the AI's prediction by changing ideally only irrelevant attributes (i.e. an altertactical image). If you move the slider below the image to the left, you will see how the image could be minimally modified to **change** the AI's prediction by changing ideally only relevant attributes (i.e. an counterfactual image).. Explore the image, you will have to interpret the AI in the following questions.



A horizontal bar with a blue slider in the center. Below the bar are three labels: 'Counterfactual' on the left, 'Original' in the center, and 'Alterfactual' on the right.

Do you think the original image is an ankle boot or a sneaker?

- ☐ ankle boot
- ☐ sneaker

Not at all
confident

Very
confident

How confident are you that your prediction is correct?

Please base your prediction in the following answer on the **explanation** shown above.

What do you think the AI will decide?

- ☐ ankle boot
- ☐ sneaker

Not at all
confident

Very
confident

How confident are you that you have predicted the decision of the AI correctly?

Next

Figure 19: Screenshot of the user study, part 9.

Prediction 1/8

Below you see an image that was classified by the AI. To aid you in your analysis of the AI, we will provide you with an additional explanation. If you move the slider below the image to the right, you will see how the image could be maximally modified **without influencing** the AI's prediction by changing ideally only irrelevant attributes (i.e. an alterfactual image). If you move the slider below the image to the left, you will see how the image could be minimally modified to **change** the AI's prediction by changing ideally only relevant attributes (i.e. an counterfactual image).. Explore the image, you will have to interpret the AI in the following questions.



CounterfactualOriginalAlterfactual

Do you think the original image is an ankle boot or a sneaker?

- ☐ ankle boot
 ☐ sneaker

Not at all confident

Very confident

How confident are you that your prediction is correct?

☐☐☐☐☐☐☐

Please base your prediction in the following answer on the **explanation** shown above.

What do you think the AI will decide?

- ☐ ankle boot
 ☐ sneaker

Not at all confident

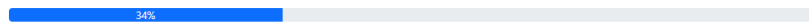
Very confident

How confident are you that you have predicted the decision of the AI correctly?

☐☐☐☐☐☐☐

Next

Figure 20: Screenshot of the user study, part 10.



Result 1/8

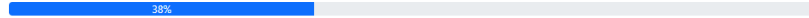
The image on the **previous** page was:



The AI predicted: **ankle boot**.

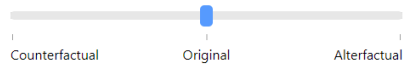
Next

Figure 21: Screenshot of the user study, part 11.



Prediction 2/8

Below you see an image that was classified by the AI. To aid you in your analysis of the AI, we will provide you with an additional explanation. If you move the slider below the image to the right, you will see how the image could be maximally modified **without influencing** the AI's prediction by changing ideally only irrelevant attributes (i.e. an alterfactual image). If you move the slider below the image to the left, you will see how the image could be minimally modified to **change** the AI's prediction by changing ideally only relevant attributes (i.e. an counterfactual image).. Explore the image, you will have to interpret the AI in the following questions.



Do you think the original image is an ankle boot or a sneaker?

- ☐ ankle boot
- ☐ sneaker

Not at all
confident

Very
confident

How confident are you that your
prediction is correct?



Please base your prediction in the following answer on the **explanation** shown above.

What do you think the AI will decide?

- ☐ ankle boot
- ☐ sneaker

Not at all
confident

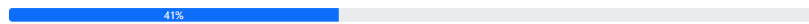
Very
confident

How confident are you that you
have predicted the decision of the
AI correctly?



Next

Figure 22: Screenshot of the user study, part 12.



Result 2/8

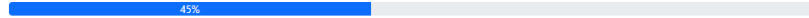
The image on the **previous** page was:



The AI predicted: **sneaker**.

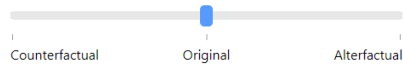
Next

Figure 23: Screenshot of the user study, part 13.



Prediction 3/8

Below you see an image that was classified by the AI. To aid you in your analysis of the AI, we will provide you with an additional explanation. If you move the slider below the image to the right, you will see how the image could be maximally modified **without influencing** the AI's prediction by changing ideally only irrelevant attributes (i.e. an alterfactual image). If you move the slider below the image to the left, you will see how the image could be minimally modified to **change** the AI's prediction by changing ideally only relevant attributes (i.e. an counterfactual image).. Explore the image, you will have to interpret the AI in the following questions.



Do you think the original image is an ankle boot or a sneaker?

- ☐ ankle boot
☐ sneaker

Not at all
confident

Very
confident

How confident are you that your
prediction is correct?



Please base your prediction in the following answer on the **explanation** shown above.

What do you think the AI will decide?

- ☐ ankle boot
☐ sneaker

Not at all
confident

Very
confident

How confident are you that you
have predicted the decision of the
AI correctly?



Next

Figure 24: Screenshot of the user study, part 14.



Result 3/8

The image on the **previous** page was:



The AI predicted: **ankle boot**.

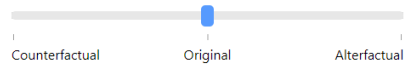
Next

Figure 25: Screenshot of the user study, part 15.



Prediction 4/8

Below you see an image that was classified by the AI. To aid you in your analysis of the AI, we will provide you with an additional explanation. If you move the slider below the image to the right, you will see how the image could be maximally modified **without influencing** the AI's prediction by changing ideally only irrelevant attributes (i.e. an alterfactual image). If you move the slider below the image to the left, you will see how the image could be minimally modified to **change** the AI's prediction by changing ideally only relevant attributes (i.e. an counterfactual image).. Explore the image, you will have to interpret the AI in the following questions.



Do you think the original image is an ankle boot or a sneaker?

- ☐ ankle boot
- ☐ sneaker

Not at all
confident

Very
confident

How confident are you that your
prediction is correct?



Please base your prediction in the following answer on the **explanation** shown above.

What do you think the AI will decide?

- ☐ ankle boot
- ☐ sneaker

Not at all
confident

Very
confident

How confident are you that you
have predicted the decision of the
AI correctly?



Next

Figure 26: Screenshot of the user study, part 16.



Result 4/8

The image on the **previous** page was:



The AI predicted: **sneaker**.

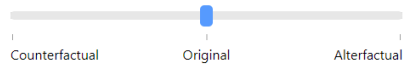
Next

Figure 27: Screenshot of the user study, part 17.



Prediction 5/8

Below you see an image that was classified by the AI. To aid you in your analysis of the AI, we will provide you with an additional explanation. If you move the slider below the image to the right, you will see how the image could be maximally modified **without influencing** the AI's prediction by changing ideally only irrelevant attributes (i.e. an alterfactual image). If you move the slider below the image to the left, you will see how the image could be minimally modified to **change** the AI's prediction by changing ideally only relevant attributes (i.e. an counterfactual image).. Explore the image, you will have to interpret the AI in the following questions.



Do you think the original image is an ankle boot or a sneaker?

- ☐ ankle boot
- ☐ sneaker

Not at all
confident

Very
confident

How confident are you that your
prediction is correct?



Please base your prediction in the following answer on the **explanation** shown above.

What do you think the AI will decide?

- ☐ ankle boot
- ☐ sneaker

Not at all
confident

Very
confident

How confident are you that you
have predicted the decision of the
AI correctly?



Next

Figure 28: Screenshot of the user study, part 18.



Result 5/8

The image on the **previous** page was:



The AI predicted: **ankle boot**.

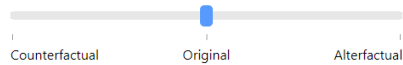
Next

Figure 29: Screenshot of the user study, part 19.



Prediction 6/8

Below you see an image that was classified by the AI. To aid you in your analysis of the AI, we will provide you with an additional explanation. If you move the slider below the image to the right, you will see how the image could be maximally modified **without influencing** the AI's prediction by changing ideally only irrelevant attributes (i.e. an alterfactual image). If you move the slider below the image to the left, you will see how the image could be minimally modified to **change** the AI's prediction by changing ideally only relevant attributes (i.e. an counterfactual image).. Explore the image, you will have to interpret the AI in the following questions.



Do you think the original image is an ankle boot or a sneaker?

- ☐ ankle boot
☐ sneaker

Not at all
confident

Very
confident

How confident are you that your
prediction is correct?



Please base your prediction in the following answer on the **explanation** shown above.

What do you think the AI will decide?

- ☐ ankle boot
☐ sneaker

Not at all
confident

Very
confident

How confident are you that you
have predicted the decision of the
AI correctly?



Next

Figure 30: Screenshot of the user study, part 20.



Result 6/8

The image on the **previous** page was:



The AI predicted: **ankle boot**.

Next

Figure 31: Screenshot of the user study, part 21.

Prediction 7/8

Below you see an image that was classified by the AI. To aid you in your analysis of the AI, we will provide you with an additional explanation. If you move the slider below the image to the right, you will see how the image could be maximally modified **without influencing** the AI's prediction by changing ideally only irrelevant attributes (i.e. an alterfactual image). If you move the slider below the image to the left, you will see how the image could be minimally modified to **change** the AI's prediction by changing ideally only relevant attributes (i.e. an counterfactual image).. Explore the image, you will have to interpret the AI in the following questions.



CounterfactualOriginalAlterfactual

Do you think the original image is an ankle boot or a sneaker?

- ☐ ankle boot
 ☐ sneaker

Not at all confident

Very confident

How confident are you that your prediction is correct?

☐
☐
☐
☐
☐
☐
☐

Please base your prediction in the following answer on the **explanation** shown above.

What do you think the AI will decide?

- ☐ ankle boot
 ☐ sneaker

Not at all confident

Very confident

How confident are you that you have predicted the decision of the AI correctly?

☐
☐
☐
☐
☐
☐
☐

Next

Figure 32: Screenshot of the user study, part 22.



Result 7/8

The image on the **previous** page was:



The AI predicted: **sneaker**.

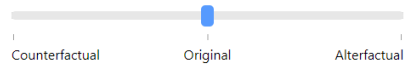
Next

Figure 33: Screenshot of the user study, part 23.



Prediction 8/8

Below you see an image that was classified by the AI. To aid you in your analysis of the AI, we will provide you with an additional explanation. If you move the slider below the image to the right, you will see how the image could be maximally modified **without influencing** the AI's prediction by changing ideally only irrelevant attributes (i.e. an alterfactual image). If you move the slider below the image to the left, you will see how the image could be minimally modified to **change** the AI's prediction by changing ideally only relevant attributes (i.e. an counterfactual image).. Explore the image, you will have to interpret the AI in the following questions.



Do you think the original image is an ankle boot or a sneaker?

- ☐ ankle boot
- ☐ sneaker

Not at all
confident

Very
confident

How confident are you that your
prediction is correct?



Please base your prediction in the following answer on the **explanation** shown above.

What do you think the AI will decide?

- ☐ ankle boot
- ☐ sneaker

Not at all
confident

Very
confident

How confident are you that you
have predicted the decision of the
AI correctly?



Next

Figure 34: Screenshot of the user study, part 24.

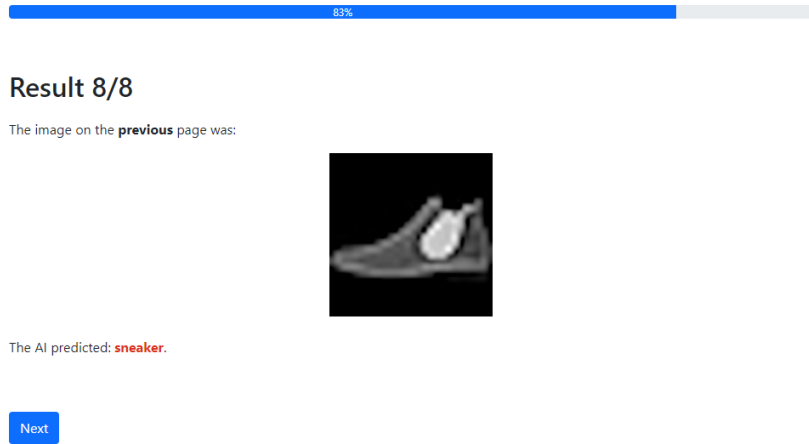


Figure 35: Screenshot of the user study, part 25.

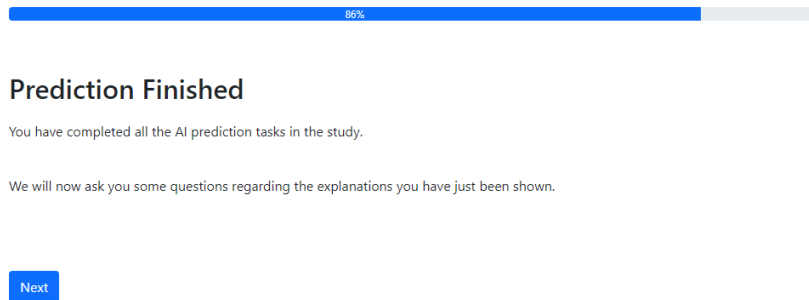



Figure 36: Screenshot of the user study, part 26.


AI Quiz

For the next question all previous images with their AI classifications are listed below. Below the images you find a slider again. By sliding it you can change the images to the corresponding explanation type like in the previous prediction tasks.


By moving the slider to the right, the image will change to a so-called alterfactual image. This alterfactual image shows you how the original image could be maximally modified **without influencing** the AI's prediction. This means that ideally only **irrelevant** attributes (for this particular decision) will be altered. By moving the slider to the left, the image will change to a so-called counterfactual image. This counterfactual image shows you how the original image could be minimally modified to **change** the AI's prediction. This means that ideally only **relevant** attributes (for this particular decision) will be altered.




AI classification:
ankle boot




AI classification:
ankle boot




AI classification:
ankle boot




AI classification:
ankle boot




AI classification:
sneaker



AI classification:
sneaker



AI classification:
sneaker



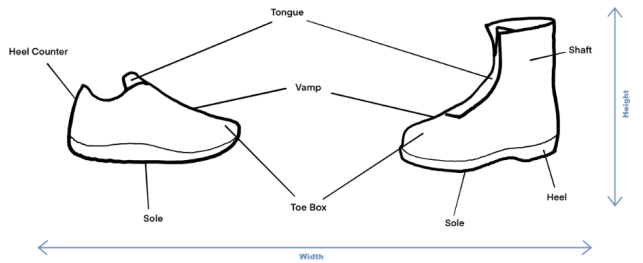
AI classification:
sneaker

Counterfactual

Original

Alterfactual

Additionally, the following terminology will be used:



In the following, you will see four statements about the AI's prediction making process. Please indicate whether you agree with the statements based on the insights you gained from the examples and explanations shown.

Note: Please base your answers solely on the examples you have been shown, **not on what you personally think should be the case.**

	Disagree	Rather disagree	I am not sure	Rather agree	Agree
The presence/absence of a boot shaft influences the AI's decision.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
If there is a boot shaft, its width influences the AI's decision.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The presence/absence of an elevated heel influences the AI's decision.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The shoe's color and pattern on the surface area influence the AI's decision.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please explain briefly why you think the AI will decide after these criteria.

Next

Figure 37: Screenshot of the user study, part 27.

93%

Explanation Satisfaction

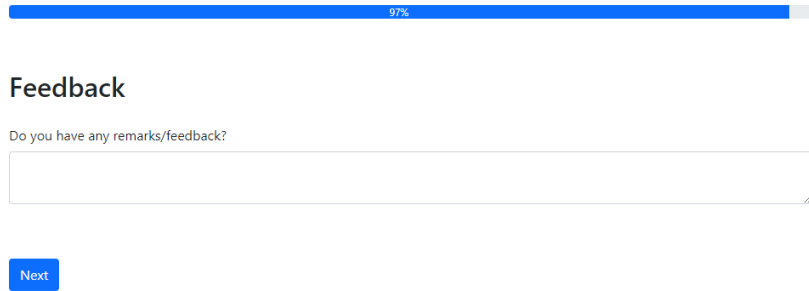
In the following statements, "explanations" refers to the alter- and counterfactual explanations, i.e. the images revealed by moving the slider below the original document.

In the following, you will see seven statements about the explanations you were shown. Please indicate whether you agree with the statements.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
From the explanations, I understand how the AI makes its decision.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The explanations of how the AI makes its decision are satisfying .	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The explanations of how the AI makes its decision have sufficient detail .	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The explanations of how the AI makes its decision seem complete .	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The explanations of how the AI makes its decision are useful to predict the AI's decision.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The explanations of how the AI makes its decision show me how accurate the AI is.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The explanations let me judge when I should trust and not trust the AI.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Next

Figure 38: Screenshot of the user study, part 28.



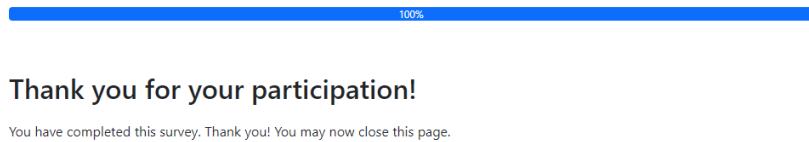
97%

Feedback

Do you have any remarks/feedback?

Next

Figure 39: Screenshot of the user study, part 29.



100%

Thank you for your participation!

You have completed this survey. Thank you! You may now close this page.

Figure 40: Screenshot of the user study, part 30.