Counterfactual Explanations & Adversarial Examples

Common Grounds, Essential Differences, and Potential Transfers

Timo Freiesleben[a,b]

[a]*Ludwigstrasse 31, Munich Center for Mathematical Philosophy, LMU, Munich, Germany*
[b]*Graduate School of Systemic Neurosciences, LMU, Munich, Germany*

Abstract

The same optimization problem underlies counterfactual explanations (CEs) and adversarial examples (AEs). While this is well known, the relationship between the two at the conceptual level remains unclear. The present paper provides exactly the missing conceptual link. We compare CEs and AEs with respect to their philosophical basis, aims, and modeling techniques. We argue that CEs are a more general object-class than AEs. In particular, we introduce the conceptual distinction between feasible and contesting CEs and show that AEs correspond to the latter.

Keywords: Counterfactual Explanation, Adversarial Example, XAI, AI-Safety, Causality

*Corresponding author
Email address: Timo.Freiesleben@campus.lmu.de (Timo Freiesleben)

1. Introduction

With the emergence of more and more flexible models in machine learning, such as deep neural networks or random forests, some new[1] problems arose. One problem is the lack of interpretability (Doshi-Velez and Kim, 2017; Rudin, 2019), which has evolved into an area called eXplainable Artificial Intelligence (XAI) or Interpretable Machine Learning (IML). A variety of model-agnostic interpretation techniques have been proposed e.g. ICE curves (Goldstein et al., 2015), LIME (Ribeiro et al., 2016), Shapley values (Štrumbelj and Kononenko, 2014). These techniques have the advantage of not posing any assumptions on the employed model (Molnar, 2019). Counterfactual Explanation (CE) (Wachter et al., 2017) is one of these model-agnostic methods and aims to explain decisions of machine learning classifiers to end-users.
Another problem with highly flexible algorithms is their vulnerability to attacks and their lack of robustness. Such an attack is called an adversarial example (AE) (Szegedy et al., 2014). Researchers constructed successful attacks (largely) for computer-vision (Goodfellow et al., 2015) but also for other tasks (Yuan et al., 2019). AEs are specific inputs that machine learning algorithms misclassify. Thereby AEs aim to deceive these algorithms and exploit their weaknesses.
Given these entirely different purposes, it is surprising that CEs and AEs share the same mathematical framework. This similarity on the model-level has been frequently noted throughout the literature. Wachter et al. (2017) describe AEs as CEs by a different name. They mention that methods are transferable but neither discuss the relationship in detail nor specify the transferable techniques. Molnar (2019) describes AEs as CEs with the aim of deception and points out the similarity as a single-objective optimization problem. Sharma et al. (2020) use counterfactuals in their robustness measure against adversarial attacks called CERScore and use the terms counterfactual/adversarial interchangeably. Tomsett et al. (2018) and Ignatiev et al. (2019) both discuss the relationship between AEs and interpretability, however, without referring to CEs. Sokol and Flach (2019) discuss CEs in the context of AI safety and note that there is "a fine line between counterfactual explanations and adversarial examples" that needs further analysis.
This paper aims to study and explicate the "fine line" between counterfactual explanations and adversarial examples. Besides a detailed mathematical analysis of the relationship between CEs and AEs, we will also conceptually compare the two and examine their common use contexts. In order to compare CEs and AEs, we need to analyze each of the two fields. For this analysis, it is important to focus on aspects that are sufficient to describe both fields (Beaney, 2018).[2] Moreover, the aspects should allow leading an informed discussion about their relationship. Hence, we have selected the following aspects:

i) conceptual basis,

ii) aim, role, and use cases

iii) models and implementations

The conceptual basis concerns the theoretical and philosophical ideas behind a concept. It describes a foundation on other, more basic ideas within the conceptual realm. The aim, role, and use cases define the motivation with a focus on the use contexts. Aspect iii) is crucial as it describes the very definition of a concept in precise mathematical terms. Moreover, it represents the state of current AI research on the topic.

[1] one might say old

[2] Note that our analysis is not a standard conceptual analysis as discussed by Carnap (1998) or Russell (1905). Here concepts are defined logically by more basic concepts. Instead, we will concentrate on a holistic picture, which also includes aspects such as the respective roles of the concepts, their use cases, state of the art, etc.

In Section 2, we point out three misconceptions of CEs and AEs present in current discussions. In Section 3, we analyze and compare CEs and AEs with respect to the aspects introduced above. In the course of this, we also discuss possible transfers between the two fields. Section 4 introduces a conceptual division of two types of CEs, namely feasible and contesting CEs. We argue that contesting CEs are similar to AEs. In Section 5 we reconsider the misconceptions discussed in Section 2 in the light of our analysis.

2. Three Misconceptions

An analysis of CEs, AEs, and their relationship on the conceptual level is urgently needed. We see several misconceptions that have already led and possibly will lead to severe confusion. Thus, our analysis's primary goal is to resolve these confusions and lay the foundation for well-guided future research on both CEs and AEs.

CE is equal to AE. The first conceptual misunderstanding is to assume that CEs and AEs are just two terms for the same objects. This misconception leads authors like Sharma et al. (2020) to use the terms counterfactual and adversarial interchangeably. The misunderstanding concerns the very basis of CEs and AEs. Not only do CEs and AEs have some non-overlapping functions, but also require AEs an additional definitional constraint that CEs do not - misclassification.
This first misconception leads to a false interpretation of robustness against attacks. Therefore, it worsens performance and societal acceptance of machine learning applications. We examine the misclassification requirement thoroughly in Section 3.1 and Section 3.3.

Feasible CEs are all relevant CEs. The second misconception appears in the context of CEs. It is the assumption that only actionable/feasible CEs are relevant to end-users. Researchers focusing on feasibility/actionability of CEs (Poyiadzi et al., 2020; Mahajan et al., 2019; Karimi et al., 2020) do not explicitly claim that. However, they neither discuss other types of CEs nor the limitations of feasible CEs. Thus, one type that is overseen are contesting CEs, which provide ground for end-users to contest a decision. Contesting CEs show a remarkable resemblance to AEs.
Focusing just on feasible CEs leads to hiding biased algorithmic decisions behind the facade of an explanation. Feasibility and contestability are discussed in Section 3.2. Section 4 introduces the distinction between feasible and contesting CEs.

Transfers, yes or no?. The third misconception is twofold. It is either over- or underestimating the transfer opportunities between CEs and AEs. In the case of overestimation, methods can be misused, or hidden assumptions can be adopted.[3] In the underestimation case, already known techniques are potentially rediscovered.[4]
Neither the adaptation of non-suited techniques nor the reinvention of successful approaches are desirable in research. To avoid these transfer problems, we will discuss conceptually permissible transfers between CEs and AEs in Section 3.3 extensively.
We will come back to these three misconceptions in Section 5.

[3] E.g Generating counterfactuals based on AE surrogate techniques (Guidotti et al., 2018) uses local approximations. Therefore, it faces the same critiques as LIME (Molnar, 2019).
[4] E.g. evolutionary algorithms for mixed data CEs (Sharma et al., 2020).

3. CEs & AEs: A Comparison

To give the reader an intuition on CEs and AEs, we start with two standard examples. The first describes a loan application scenario and a potential CE in that situation. The second example illustrates AEs in image recognition tasks.

CE Example. Assume person P wants to obtain a loan and applies for it through the bank's online portal. The portal uses an automated, algorithmic decision system, which decides that Person P will not receive the loan. P wants an explanation for that decision. An example of a CE would be:

> If P had a higher salary and an outstanding loan less, her loan application would have been accepted.

AE example. Look at Figure 1 from Papernot et al. (2017). The images of row two are a slight modification of the images from row one. However, row two shows AEs since these subtle changes have changed the classification to the wrong classes. The image recognition algorithm was successfully tricked.

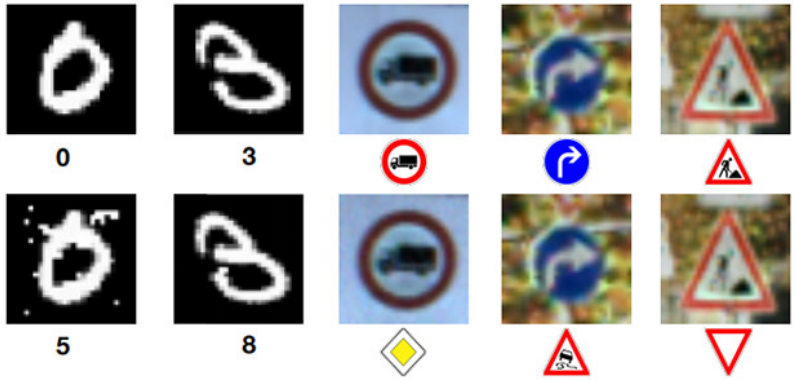

Figure 1: In the first row, we can see five images (The first two are from the MNIST dataset, the other three are from the GTSRD dataset.) that are classified correctly. We see the same five pictures in the row below but slightly modified by some noise added to the pictures. Here, the algorithm misclassifies them.

3.1. Basis

The first aspect we investigate is the conceptual basis of CEs and AEs. We start by considering each of them in separation, and then we conduct our comparison.

Counterfactual Explanations

CEs have a strong philosophical basis and tradition. Here, we confine ourselves to counterfactuals in the form of subjunctive[5] conditionals. Let S and Q be propositions. Then, *counterfactual sentences* are conditionals of the form:

$$\text{If } S \text{ was true } Q \text{ would have been true.} \tag{1}$$

[5]The difference between indicative and subjunctive conditionals is that the antecedent must be false for the latter (Starr, 2019). From now on, whenever we talk about counterfactual statements/sentences/explanations/conditionals we mean subjunctive ones.

Importantly the antecedent of the conditional, namely S is false. A *counterfactual explanation* is a counterfactual sentence that is true. What makes counterfactual statements true is hotly debated in philosophy, and no solution has been agreed upon generally (Starr, 2019). The solution taken up in the computer science approach builds on the work of Lewis (1973). In Lewis's framework, Equation (1) holds if and only if the closest possible world $\omega' \in \Omega$ to the actual world $\omega \in \Omega$ in which S is true[6] also Q is true.[7] Due to the under-specified notion of similarity between possible worlds, Lewis's proposal is highly controversial (Starr, 2019).

A *good counterfactual explanation* in a specific situation is a CE relevant to the explainee. That means that the CE is easy to comprehend and has an interesting[8] antecedent/consequent for the explainee (Miller, 2019). In XAI applications, the antecedent describes a change in features from a given input, and the consequent describes a change in the outcome of the classification.

Note that Lewis aimed to describe causality via counterfactuals (Menzies and Beebee, 2019), which is not directly the goal of CEs in XAI.[9] The CE approach allows us to make causal claims about the machine learning model only (e.g., which features does the algorithm take to be causally relevant) but not the corresponding real-world objects (Molnar et al., 2020).

Adversarial Examples

Adversarial examples are inputs that an algorithm assigns the *wrong class/value* to.[10] A wrong class is defined by the fact that the class deviates from a ground truth given by humans. Not for all inputs, there are such ground truths. Especially for large feature spaces, there are many entirely meaningless inputs, which are not considered AEs. What defines AEs is that the given inputs appear similar (or identical) to real-world data or algorithm training data. Generally, this is achieved by modifying a real-world input. *Good AEs* are those that can potentially be exploited by an attacker.

AEs base on a classical scenario from game theory, in which a deceiver-agent tries to trick a discriminator-agent (Dalvi et al., 2004).[11] An AE denotes a successful deception in this game. Adversarial attacks are not specific to neural networks but apply to any complex system (Papernot et al., 2016a); not even humans are immune against deception (Kahneman et al., 1982; Chabris and Simons, 2010; Ioannou et al., 2015). Remarkable about modern AEs is that they often transfer from one model to another and are hard to get rid of (Yuan et al., 2019). The origin of this effect is still open to debate (Goodfellow et al., 2015; Ilyas et al., 2019).

Comparison

Both fields rely on counterfactual reasoning. CEs describe a variation of the actual situation/world; AEs are a real-world input variation. Also, in both cases, the variation changes the result. Furthermore, both approaches search for relevant variations in regions close to the real world (input), fulfilling certain constraints. The crucial difference lies in these constraints. For CEs as given in Equation (1), the alteration to the actual world described by a predicate S has to satisfy that predicate Q applies. For AEs, an alternative to

[6] As mentioned, above S is false in ω.

[7] Note that Ω denotes the set of possible worlds.

[8] Section 3.2 specifies this intuitive notion of interestingness.

[9] Contrary to Pearl (2009) who introduces CEs with causal meaning. An XAI version of this type of CEs presents Karimi et al. (2020) in the form of algorithmic recourse. Moreover, counterfactuals can also account for non-causal explanations, as discussed in Reutlinger (2018).

[10] From now on, we will mainly talk about misclassification and classifying. However, this is only to simplify our language usage. AEs are not restricted to classification tasks but also work on regression problems.

[11] This picture of the two opponents is also the basis of generative adversarial networks (Radford et al., 2016).

a real-world input must be misclassified compared to some ground truth. If we define Q as the predicate 'Misclassified', we see that the latter is a specific case of the former. Misclassification is not demanded of CEs. Conversely, CEs often demand that Q describes a specific outcome, such as 'loan acceptance'.

3.2. Aim, Role, and Use Cases

Now, we investigate the aim, role, and use cases of CEs and AEs.

Counterfactual Explanations

The CE approach in XAI aims to generate local explanations. Here, local means that the explanations are generated for individual "decisions"[12] of the algorithm. According to Wachter et al. (2017) and Miller (2019), these explanations have three intuitive aims, which make the difference between (only) a CE and a good CE. The aims are to

i) raise understanding[13],

ii) give guidance for future actions, and

iii) allow to contest decisions.

A good CE does not have to meet all three of these goals. In many contexts, explanations focus on only one or two of them.

Aim i). The target audience of CEs are laypersons who are neither experts in machine learning nor have unlimited time resources (Wheeler, 2020). If we want to improve a person's understanding, we must respect these resource limitations and focus on the few main reasons for a decision. To achieve this degree of simplicity, we must be economical concerning the presented number of reasons. Therefore, *sparsity* is one aim discussed in the literature.

Aim ii). Explanations should serve as a guideline for future actions. Hence, the alternative that achieves the desired output (e.g., obtaining the loan) should be reachable for the explainee. For example, a loan applicant cannot become younger to obtain a loan, even if age is one of the bank's criteria for justified reasons. Thus, it is not reasonable to propose a reduction in age for obtaining a loan. Researchers summarize such limitations under the term *feasibility*.

Aim iii). Explanations provide grounds for the appealability of decisions. End-users can contest a decision if the presented reasons are insufficient or discriminatory. That may be because the decision is based on features that should not play a role (e.g., skin color, gender, etc.) or on features that should play a role but are expected to have a different effect (e.g., if a high salary correlates negatively with obtaining a loan). All in all, humans want to be treated fairly and demand explanations to uncover unfair judgments (Kusner et al., 2017; Asher et al., 2020). If we feel unfairly judged, this is because we would have expected a different decision. Thus, the explanation we generate should focus on features that the explainee expected to have a different effect on the decision. In other words, explanations should be informative.[14]

[12] By "decisions", we usually mean classification or regression tasks the algorithm performs.

[13] Páez (2019) argues that counterfactuals as introduced by Wachter et al. (2017) can even in principle not meet this requirement.

[14] The condition of *informativeness* also aids aim number one, which is to raise understanding. In the context of psychology, informativeness is discussed under the name abnormality (Miller, 2019).

Role. Among XAI researchers, CEs became very popular. One reason is that they are model-agnostic and, therefore, applicable to any kind of algorithm. Secondly, there is a one-to-one correspondence to contrastive explanations, which are the type of explanations that people use most in everyday life (Miller, 2019). Thirdly, CEs are compatible with the right to explanation in the European General Data Protection Regulation (GDPR) (Wachter et al., 2017). All these advantages allow the CE approach to playing an essential role in XAI.

Use Cases. The use cases of CEs are generally unlimited as the CE framework applies to any kind of algorithm. The only requirements are interpretable input- and output-spaces and reasonable distance measures on these spaces. In the literature, however, CEs are considered exclusively in connection with classification tasks on tabular data. Unsupervised/Reinforcement learning, image/audio classification, or regression problems are still under-explored in this respect.[15] Whether CEs should be applied on a broad scale is at least questioned (Laugel et al., 2019; Barocas et al., 2020).

Adversarial Examples

Even though AEs represent only single instances[16] in which the algorithm fails, they also point to the algorithm's global problems. If the algorithm classifies a stop sign as a right-of-way sign, one becomes extra cautious about everything the algorithm does. The aims behind AEs depend strongly on the specific use cases and employers. Nevertheless, we find three principal aims that all perspectives share:

i) to fool the system,

ii) to do this imperceptibly. and

iii) effectively.

What imply these aims for generating AEs?

Aim i). Fooling the system is the main aim of AEs. Thus, we look for *missclassifications*. The system usually performs reasonably well on training data and similar inputs. In unseen regions, on the other hand, it performs poorly. If we randomly pick an example from unseen regions of our input space, we most likely choose a meaningless data point.[17] So we need to find an input that is close enough to a meaningful input and yet in a region where the algorithm performs poorly.

Aim ii). One condition under which we must search for adversarials is *imperceptibility*. AEs should not be easy to detect for a human, i.e. input changes should be below the human perception threshold. This property guarantees the highest chance of deceiving successfully. Since human perception directs attention to certain features and expectations, imperceptibility can be achieved in two ways. Either by changes in unattended features or by distributed low-intensity changes.

Aim iii). The effectiveness of an AE depends strongly on the context. Attackers want to exploit mistakes in the most profitable way possible (e.g., monetary gain or system damage). Engineers want to defend themselves against such attacks and make their system more stable against them (e.g., fixing bugs or detecting attacks). Researchers working on AEs strive for a deeper understanding of learning algorithms, depicting real-world dangers in employing algorithms, and high research impact.

[15] One counterexample to this is the work on the MNIST dataset by Van Looveren and Klaise (2019).
[16] They can also be aiming at a global level and a variety of algorithms such as shown by Moosavi-Dezfooli et al. (2017).
[17] Especially considering images

Role. AEs are both a blessing and a curse. They can indeed cause significant harm to individuals, companies, and society as a whole. The more social or ethical consequences the task we assign to a machine learning algorithm has, the worse the effect of misclassification. A stop sign classified as a right of way sign can cause accidents, and a rifle misclassified as a turtle can facilitate terrorist attacks at airports. The trust we have in AI systems is and will be closely linked to the extent to which adversarial attacks are possible. On the positive side, AEs can help us understand how the algorithm works (Ignatiev et al., 2019; Tomsett et al., 2018). Knowing where the algorithm has problems helps us understand what the algorithm is really learning (Lu et al., 2017). Moreover, by adversarial training, AEs can even concretely improve models (Bekoulis et al., 2018; Stutz et al., 2019).

Use Cases. AEs are mostly built for image and sometimes audio recognition tasks.[18] Reasons for that are uncontroversial ground-truths, the boom in computer vision, and the resemblance with optical illusions (Elsayed et al., 2018).

Comparison

Both fields help to understand what algorithms have learned. Moreover, both contribute to identifying biases and even offer methods to eliminate these biases through adversarial- or counterfactual-training (Bekoulis et al., 2018; Sharma et al., 2020). However, while improving understanding and highlighting algorithmic problems is usually only a byproduct of AEs, it is the focus of CEs. The deception of a system, on the other hand, is essential for AEs, but a potential byproduct of CEs in cases where they disclose too much information about the algorithm (Sokol and Flach, 2019).[19]

Making modifications imperceptible is crucial for AEs. In the case of CEs, however, the modifications form the core of the given explanation. This is more a difference in presentation and less one in the type of modifications. Modifications to achieve the imperceptibility of AEs show a great similarity with those in CEs that aim at informativeness. Imperceptible AEs result from modifications in unnoticed/unanticipated but effective features. These surprisingly effective changes are precisely those, which are most informative to humans, as Section 4 discusses.

For the feature permutations in CEs to make sense, the input space must provide a certain degree of interpretability. This demand is irrelevant for AEs since the changes are hidden and not highlighted.[20]

The two approaches play a similar role within the machine learning landscape, as both will strongly affect people's trust in machine learning systems in the future. Also, both fields have gained increasing legal relevance. To be legally applicable (e.g., in autonomous driving, airport security (Athalye et al., 2018), etc.), machine learning algorithms must be both robust against AEs and provide explanations to end-users as specified in the GDPR. A significant difference is that AEs, by definition, can only point to the mistakes of the algorithm. Hence, emerging AEs mainly have a negative role, while CEs can also raise trust in the system.

Concerning the use cases, we see that AEs mainly apply to computer vision tasks, whereas CEs are considered only on tabular data.[21]

[18] Ballet et al. (2019) give an example of AEs for tabular data classification.

[19] Meaning that giving guidance for future actions and deceiving are only compatible for immoral agents.

[20] One difference is that CEs only tell us how an algorithm works in a very local region, while AEs can affect humans' confidence in the system as a whole. However, in cases where the CEs reveal racist, sexist, or causally unjustified reasons, this will also reduce humans' confidence in the system as a whole, not just locally.

[21] Even though there are the above-mentioned counterexamples to that division (Van Looveren and Klaise, 2019; Ballet et al., 2019).

3.3. Models and Implementations

The last conceptual aspect we investigate concerns the models and implementations of CEs and AEs.

Counterfactual Explanations

There are a variety of formulations of the CE framework and the present version orients at Wachter et al. (2017). Assume there is a *learning algorithm*[22], which we represent by a function $f : I \to O$ mapping a vector x from an *interpretable* (potentially high dimensional) input space I to a vector y in an output space O. Assume the desired classification for x would be $y' \neq y$. Then, a *counterfactual vector* $x_{c,y'}$ to x is a vector that minimizes the term $\|x_{c,y'} - x\|$ for which $f(x_{c,y'}) = y'$. Often it is sufficient that $x_{c,y'}$ is close to x. Also, having $f(x_{c,y'})$ close to y' can be sufficient as it might be in principle impossible or very difficult to reach. Thus, the standard formulation as a single-objective optimization problem is

$$\underset{x' \in I}{\operatorname{argmin}} \|x - x'\| + \lambda \left\|\!\left| f(x') - y' \right|\!\right| \tag{2}$$

where $\|\cdot\|$ and $\|\!|\cdot|\!\|$ are induced by some measures of distance on I and O respectively.[23] The scalar λ trades off between a more similar counterfactual and a vector closer to the desired output. The *counterfactual explanation* is derived by the difference between the original input and the counterfactual vector we generated put into words.

Consider, for instance, the loan application scenario 3 and assume that $x_{c,y'} - x$ has a value of +2000€ at the feature salary and a value of -1 at the feature open loans. Then, the corresponding CE would be:

- If P earned 2000€ more per year and had one outstanding loan less, her loan application would have been accepted.

Distance Measures. As in Lewis's framework from Section 3.1, the main difficulty is to define a reasonable distance measure on the input space[24]. As discussed in Section 3.2, good CEs are sparse, feasible, and informative.

Sparsity is a hot topic in the general machine learning literature (Bach, 2010). In the field of CE, Wachter et al. (2017) gain sparsity by using the normalized Manhattan metric. Other ways to attain sparsity include setting features as not permutable (Moore et al., 2019), using the L_0 metric to directly penalize high numbers of changed features, using multi-objective optimization with the number of changed features as one objective (Dandl et al., 2020), or taking into account the causal structure of the real world where changing a few features via an action has consequences for several others (Karimi et al., 2020).[25]

Feasibility can be achieved in many ways and depends on the problem under consideration. One possibility is to declare some immutable features, making them irrelevant for the explanation (Moore et al., 2019; Sokol and Flach, 2019). The second way focuses on the problem that some possible input vectors represent highly improbable or unreachable combinations of features in the real world. They should therefore, not be suggested as good CEs. The literature suggests many ways to deal with this problem. Examples are

[22] Usually this algorithm is already trained.

[23] They don not necessarily have to be norms.

[24] Not only on the input space, it might be challenging to find a suitable measure of distance but also on the output space. Consider a classification problem, where the output space is not just a set of options but a set of probability density functions. If the desired output is "loan application accepted" it is unclear whether it has the highest value among the categories, more than fifty percent, or even the value 100%. Moreover, some outcomes might be more similar to the desired outcome than others, e.g., obtaining a smaller loan is better than obtaining no loan. Standard measures like KL-divergence or cross-entropy are ignorant to such similarity differences.

[25] Moore et al. (2019) also introduce the idea to show a range of explanations with a diverse number of changed features.

considering the probability density of inputs (Sharma et al., 2020; Kanamori et al., 2020), the distance to the training data (Dandl et al., 2020), the causal structure of the real world (Karimi et al., 2020; Mahajan et al., 2019), or the lengths of the paths between the original input and the counterfactual (Poyiadzi et al., 2020). While sparsity and feasibility are discussed throughout the literature, informativeness has not been considered. For informativeness, we demand information about the explainee's estimates/expectations, which is usually not available. However, some solved this problem by asking the user questions about her preferences (Sokol and Flach, 2019). Another option is to focus on the features that the average human usually over- or underestimates. Best would be a combination of the two, i.e., to set the average human's prior characteristics and then update this prior via feedback from the human agent.

Solution Methods. The solution strategy for the optimization problem depends on the model-knowledge. As the employers of interpretation techniques are usually the designers of the inspected algorithm, full model access is common. Given such a white box, the problem can be solved by gradient-based methods (Wachter et al., 2017; Mothilal et al., 2020; Mahajan et al., 2019). An alternative for mixed numeric/categorical data are mixed-integer linear program solvers (Ustun et al., 2019; Russell, 2019; Kanamori et al., 2020). Genetic algorithms are a solution method that does not require model knowledge (Sharma et al., 2020; Dandl et al., 2020). A more controversial technique for black-box scenarios is to train a surrogate model on the original model and then transfer the CEs from the surrogate to the original model Guidotti et al. (2018).[26]

Selection Problem. The solution to the optimization problem will generally not be unique. There can be a high number of equally close CEs for the same input vector. Worse, these different CEs may provide explanations that are pairwise incompatible as the following two:

- If P earned 2000€ more per year, her loan application would have been accepted.
- If P earned 2000€ less per year, her loan application would have been accepted.

Such cases arise since the decision boundaries do not follow classical monotonicity constraints. For example, the state may subsidize loan applications from people below a certain salary level. Some propose therefore to present several different CEs like Mothilal et al. (2020); Moore et al. (2019); Wachter et al. (2017); Dandl et al. (2020). However, then the question arises, how many and which ones? Others propose to select a certain CE according to relevance (Fernández-Loría et al., 2020) or a quality standard set by the user, such as complexity or particularly interesting features (Sokol and Flach, 2019). The question remains open as to how this so-called Rashomon effect can be solved.

Adversarial Examples

There are a variety of formulations of the AE framework. The version presented here orients at Yuan et al. (2019). Since the framework is basically the same as for CEs, we will mainly focus on its deviations. Again, the learning algorithm is represented by a function $f : I \rightarrow O$ mapping a vector x from an input space I to a vector y in an output space O. AEs do not require an interpretable input space. We distinguish between a targeted and a non-targeted attack. For a *targeted attack*, a particular alternative output $y' \neq y$ is desired, as in the case of CEs. For a *non-targeted attack*, the alternative output just has to differ from y. For a non-targeted attack, an AE x_a to x is generated by searching for an x_a that minimizes $\|x_a - x\|$ and for which $f(x_a) \neq f(x)$. In case we do have a particular alternative output in mind, the adversarial $x_{a,y'}$ to x is a vector

[26] Problems occur if the surrogate model is not faithful to the original model. In such cases, the CEs generated are simply false and potentially misleading.

that minimizes the term $\|x_{a,y'} - x\|$ for which $f(x_{a,y'}) = y'$. Equation (2) presents one formulation as a single objective optimization problem. As in the case of CEs, the minimality might not be as important, and it is enough to find inputs close enough to x that change the classification in the desired way. Considering more distal inputs might be computationally easier and more interesting in some cases (Elsayed et al., 2018).
Of major importance is that the input is misclassified, which is guaranteed by none of the above-mentioned optimization problems. To achieve misclassification, we must add the condition that the alternative input is incorrectly classified.[27] In other words, for the adversarial x_a, respectively $x_{a,y'}$ has to hold $f(x_a) \neq y_{true}$ respectively $f(x_{a,y'}) \neq y_{true}$. Here, y_{true} denotes the actually correct label for the adversarial example. This true label is usually the same as for the original input x, namely y.
The optimization problem presented here is only one among many (Yuan et al., 2019). Also, formulating an optimization problem is not the only way of finding AEs. The fast gradient sign method of Goodfellow et al. (2015) is an example of how to generate AEs directly.

Distance Measures. One of the most critical problems in creating an AE is the reasonable definition of a distance measure on the input space - both computational and qualitative aspects have to be considered. This leads us to the aims of misclassification, imperceptibility, and effectiveness.
Again, minimizing the difference between x and its adversarial x_a and flipping the algorithms assignment does not guarantee to attain an AE. A switch in classification due to a small variation may be justified.[28] However, since we are often dealing with image data, a tiny variation[29] rarely justifies a switch in classification. It, therefore, makes it an AE. If we look at image data, there are usually infinitely many meaningless data points between two proper classes. Hence, following the gradient (Goodfellow et al., 2015), the Jacobian (Papernot et al., 2016b) or any other reasonable procedure (Yuan et al., 2019) may easily lead to an AE. Section 4 gives further insights into the problem of misclassification.
Imperceptibility is realized in various ways in the literature. Some change very few or even one feature strongly by optimizing for the L_0 norm (Su et al., 2019). Others alter more features to a smaller amount with the L_1 norm (Carlini et al., 2018), or in a variety of real-world contexts and scenarios, like Brown et al. (2017); Athalye et al. (2018). The standard way to gain imperceptibility is to alter all features slightly via the L_∞ norm on the input space (Goodfellow et al., 2015; Szegedy et al., 2014). Basically, any p-norm can be reasonably applied (Yuan et al., 2019). More interesting are measures that take into account what humans consider as "close" inputs (Rozsa et al., 2016; Athalye et al., 2018). This approach leads to an overlap between human and machine deceivability (Elsayed et al., 2018). For tabular data classification, it is much harder to define imperceptibility. Ballet et al. (2019) solved this via defining critical and non-critical features[30]. Since the algorithm uses both types of features in the classification, they modify only non-critical features to attain a change in assignment. This example shows how imperceptibility and misclassification go hand in hand.
Effectiveness is not so much a question of defining the distance measure but rather a question of which example we use to build our AE.

Solution Methods. The community's main focus is on the algorithmic generation of AEs, which again differs depending on model knowledge. There are gradient-based methods for white boxes, either for solving

[27] In the case of a regression problem, this could correspond to being far outside the range of reasonable output values, see Balda et al. (2019).
[28] Consider a case, where a loan application of a 17-year-old is rejected while an 18-year-old with the same characteristics would receive the loan.
[29] Often this variation is not only small in the sense of a p-norm, it is moreover structureless noise.
[30] Based on expert evaluation

the optimization problem (Szegedy et al., 2014; Athalye et al., 2018; Brown et al., 2017) or for the direct generation of AEs (Goodfellow et al., 2015). Other options include the Jacobian (Papernot et al., 2016b) or neural network feature representations (Sabour et al., 2016). In addition to white-box attacks, there are many black-box solution methods, such as the approximation of gradients via symmetric differences (Chen et al., 2017) or evolutionary algorithms (Guo et al., 2019; Alzantot et al., 2019; Su et al., 2019). Due to the transferability of AEs, it is often also possible to build an AE for a surrogate model and then apply the AE to the original model (Papernot et al., 2017). Generally, non-targeted attacks are computationally less costly and do more easily transfer to other systems than targeted attacks (Yuan et al., 2019).

Selection Problems. If we want to generate an AE, we face two selection problems. Probably due to their irrelevance in practice, neither of them has been discussed in the literature. The *first selection problem* is about selecting the initial input vector the AE is based on. The *second selection problem* is about selecting the final AE among the solutions to the optimization problem. Both selection problems relate to effectiveness and depend on the application, the goal, and the system's weaknesses. The AEs that best suit the employer's needs should be picked.

Comparison

There is no need for researchers to reinvent the wheel. For this reason, this section discusses, in addition to the conceptual comparison, the potential transfers between the fields Figure 2.
The common ground with regard to the mathematical model is evident. In Appendix A we show that AEs are special solutions to a (non-targeted) CE optimization problem.

Theorem (Every (targeted ϵ) AE is a (targeted ϵ) CE). *For all $x \in I$, $f(x) \neq y' \in O$, $\epsilon > 0$ and distance measures $\|\cdot\|$ holds:*

i) $AE(x, \epsilon, \|\cdot\|) \subseteq CE(x, \epsilon, \|\cdot\|)$

ii) $TAE(x, y', \epsilon, \|\cdot\|) \subseteq TCE(x, y', \epsilon, \|\cdot\|)$

iii) $AE(x, \|\cdot\|) \subseteq CE(x, \|\cdot\|)$

iv) $TAE(x, y', \|\cdot\|) \subseteq TCE(x, y', \|\cdot\|)$

This means that (targeted) AEs are (targeted) CEs that are misclassified. As we show, this holds also for AEs and CEs in a given ϵ-environment. However, it is important to point out that non-targeted attacks are common while non-targeted counterfactuals are rather rare. Moreover, some formulations as optimization problems already encode the respective aims as in the case of e.g. Dandl et al. (2020); Van Looveren and Klaise (2019) for CEs and e.g. Carlini and Wagner (2017) for AEs. For formulations that are targeted and where the respective aims are not encoded in the optimization problem, transfers between the fields are permissible. Interestingly, we do not necessarily need to formulate an optimization problem to generate AEs (Goodfellow et al., 2015). Direct generation methods are theoretically also possible for counterfactuals, even though the generated CEs will be much harder to justify conceptually.
The different aims are mostly not encoded in the optimization problem but the distance measure. For that reason, we find the most significant differences between the fields if we consider the distance measures. However, there are also similarities to be found. Notions of distance that realize sparsity show commonalities with those that realize imperceptibility. A change in few among lots of features is often difficult to spot (Su et al., 2019), especially when the change is not highlighted. Distance measures that favor sparsity can, therefore, be desirable to transfer between the fields. Moreover, distributed changes to achieve the imperceptibility of AEs in e.g., images are not per se irrelevant for CEs. The sparsity of CEs is only relevant for

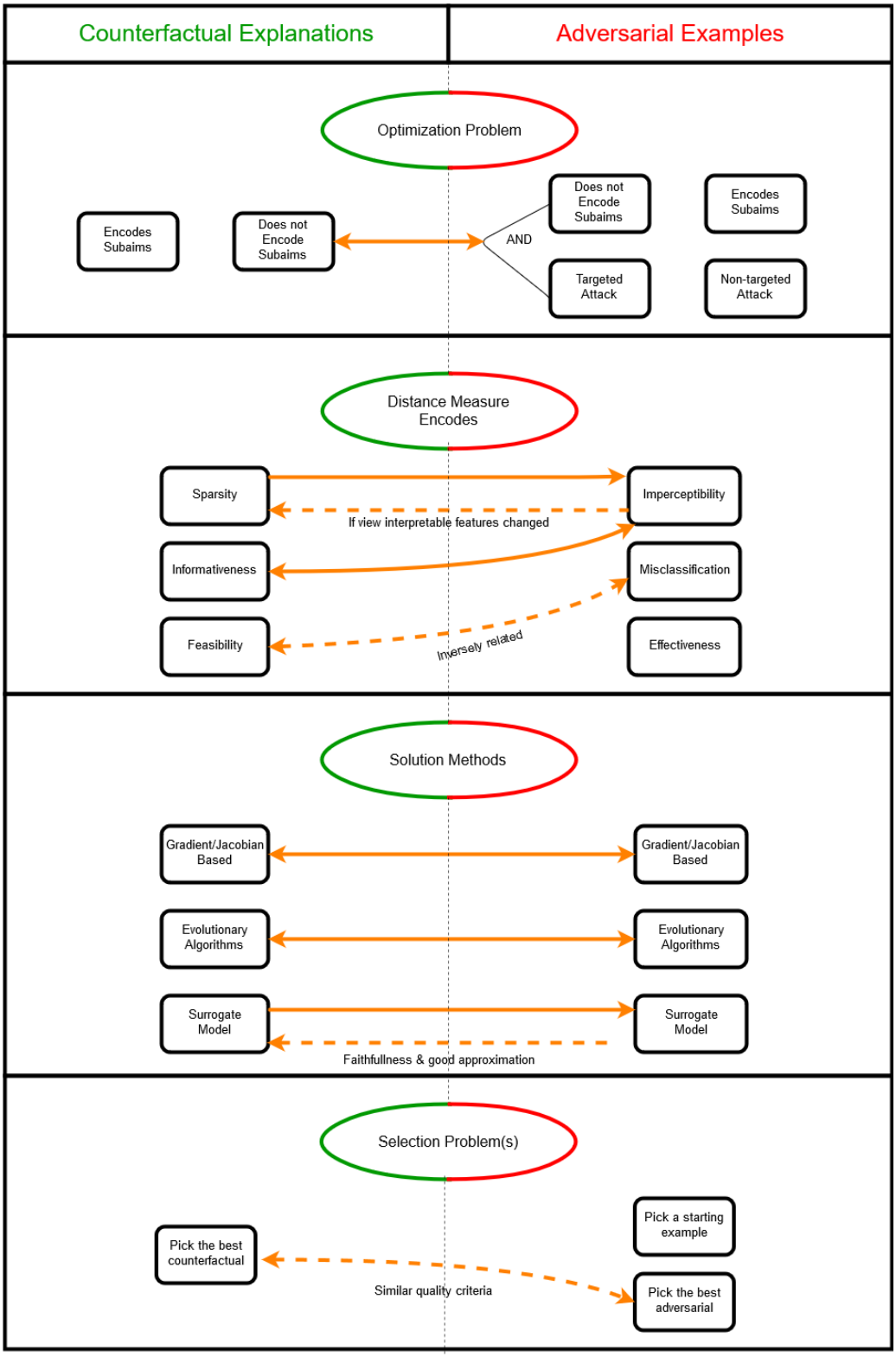

Figure 2: On the left-hand side, there is the counterfactual realm and on the right-hand side the corresponding adversarial concepts. Solid arrows between two items mean that a transfer is allowed in that direction. Dashed arrows mean that a transfer is possible under additional conditions, specified below the arrows.

changes in interpretable features. For non-interpretable features, distributed changes can often be described as sparse changes in more abstract interpretable features.
The CE aim of informativeness and the AE aim of imperceptibility also align and Section 4 discusses this in further detail. Changing unexpected but effective features will often lead to imperceptible changes. As people assume these features to be non-effective, they pay little attention to them. The same holds vice versa. Hence, we can expect fruitful transfers. Feasibility counteracts the goal of misclassification Feasibility in CEs requires that the explainee can realistically reach the alternative data point generated. Realistic datapoints are those that are generally well represented in the training data. Hence the algorithm usually performs well in such cases. However, it may be possible to reverse distance measures that encode feasibility to aid in misclassification. Furthermore, there are cases where feasibility can be relevant for AEs, such as anomaly detection or the generation of realistic AEs.
Due to their similarity in the optimization problem, the two approaches also use similar solution methods. We can observe this parallelism in the development of the fields. Both started with gradient-based methods, proceeded with evolutionary algorithms, and then considered surrogate models.[31] If applicable, solution methods developed for CEs are also suited to generate AEs. The opposite direction might be more problematic. For CEs, approximately good solutions are often not good enough because they lead to bad/misleading explanations. This problem becomes particularly apparent when we look at the, among AE researchers, highly popular surrogate model approaches. If the surrogate model is not faithful enough to the original model, the generated CEs will end up being wrong and, in the worst-case, misleading.
It is already noteworthy that both fields face selection problems. Moreover, in both approaches, CEs/AEs among all the solutions to the optimization problem must be selected. Nevertheless, the differences prevail. First, for AEs, an initial input has to be selected, whereas, for CEs, this input is given by the end-user. Second, First, the solution space to non-targeted AEs contains vectors from different classes, not one as for CEs. Third, the number of presentable CEs is limited by humans' capacity to process information, while the number of AEs to try out is unlimited.

4. The Two Types of CEs

Until now, we have discussed the similarities, differences, and possible transfers between CEs and AEs. However, we left out the relationship at the level of individual instances. Does a good CE make a good AE or vice versa? In this section, we present a conceptual division into two types of CEs. We call them *feasible CEs* and *contesting CEs*. Feasible CEs are reasonable explanations that allow for deriving future actions. Contesting CEs are explanations that provide a basis for challenging an automated decision. AEs link closely to the latter. We believe that defining and analyzing these two types clarifies existing misunderstandings regarding the relationship between CEs and AEs. In our analysis, we will look at several real-world scenarios.
For all the presented scenarios, we presuppose the following:

- There is a trained supervised-learning algorithm represented by a function $f : I \rightarrow O$ mapping a vector x from a (potentially high dimensional) input space I to a vector y in an output space O.

[31] Interestingly, the areas concentrate on different solution methods. While the literature on AEs mainly discusses white box solvers, the literature on CEs focuses on black-box solvers. This observation is surprising since AEs are usually considered from an attacker's perspective without access to the model, while CEs are generated by the model engineers. A look at the use cases explains this paradox. If we consider low-dimensional tabular data and standard algorithms, simple black box attacks are perfectly feasible. For high dimensional image data and deep convolutional neural networks, on the other hand, black-box attacks explode computationally.

- There is a true causal graph C describing the causal relationship between the set of input features $\mathcal{I}$ and the set of output features $\mathcal{O}$. Each of the input and output features relates to real-world items, and C describes these items' true causal structure. Often, relevant features to complete the full causal picture from reality are missing, which can be either compensated for via latent variables Pearl (2009) or is ignored[32]. Since ignoring them is common practice in supervised-learning contexts, we proceed with the latter. We will call $F \in \mathcal{I}$ a *causally relevant feature* for $T \in \mathcal{O}$ if either
  - F is an ancestor node of T in the causal graph C or
  - there is a common cause $L \notin \mathcal{I} \cup \mathcal{O}$ of both F and T that is not part of the causal graph.[33]

  We say $F \in \mathcal{I}$ is a causally irrelevant feature for $T \in \mathcal{O}$ if neither of the two conditions is met.
- In decision making, humans pay most attention to variables they consider relevant for the task (Jehee et al., 2011; Ballet et al., 2019). Well-trained decision-makers, therefore, focus on causally relevant variables (Navalpakkam and Itti, 2005). Hence, they often oversee changes in causally irrelevant features, which makes these changes imperceptible.
- For our example scenarios, again consider loan applications. For simplicity, we make the unrealistic assumption that the input space I only contains information about the features salary and the number of dogs. The output space O is a binary feature that either takes the value 1 for loan acceptance and 0 for loan denial. We assume that Figure 3 expresses the real causal relationship of the involved variables. That means that the number of dogs should be irrelevant for loan approval given we know the salary. A high salary is a good reason for loan acceptance and a necessary condition for having many dogs (generally expensive). Thereby, the features number of dogs and loan approval are correlated. This causal graph will help us in depicting the relation between feasible and contesting CEs in different scenarios. The setting is inspired by Ballet et al. (2019) who built AEs for tabular data.

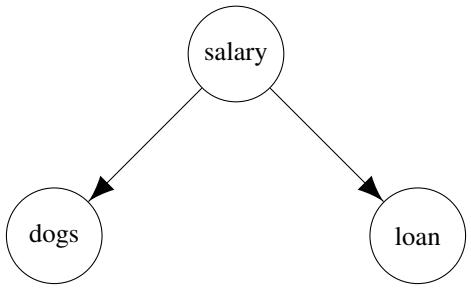

Figure 3: The causal graph contains three variables, salary, the number of dogs, and a binary variable for the loan application status.

Both types of CEs we introduce here indicate the features the algorithm finds relevant for the decision process. However, they differ in the kind of features they change. Feasible CEs permute causally relevant features. Contesting CEs, on the other side, point out which causally irrelevant features played a role in the decision process. We also discuss mixed type CEs with both causally relevant and irrelevant features permuted.

[32] Some features might even be inaccessible.

[33] Causal relevance in the first condition is clear. In the second condition, F is only indirectly causally relevant for T. F gives us information about the unknown variable L which is a cause of T.

4.1. Feasible CEs

To make the division maximally clear, we consider a scenario where only feasible CEs exist. These cases occur in the presence of perfect algorithms. A *perfect algorithm* describes a case where the algorithm's decisions match the ground truth in all scenarios where such a ground truth exists. That is, if we consider an input $x \in I$ then $f(x)$ is correctly classified. In this case, no AE exists since misclassification is a necessary condition for an AE. CEs, on the other side, do exist.[34]

Scenario. Assume the algorithm f is perfect. Thus, for any combination of the number of dogs and the salary for which a ground truth exists, the algorithm maps exactly to that ground truth. Hence, the algorithm learned that all that is relevant for loan acceptance is the salary. Given the salary surpasses a certain threshold t, the algorithm grants the loan. If the applicant has a salary s below t and a given number of dogs d, the counterfactual vector would be (t, d). The corresponding CE would be:

- If P's salary was $(t - s)$€ higher, her loan application would have been accepted.

This explanation would indeed be good because it aids understanding and guides future actions. These are precisely the functions of feasible CEs - they permute causally relevant properties to the right amount.

4.2. Contesting CEs

AEs do not exist for perfect algorithms. As we will see later, contesting CEs are just like AEs. Thus, we need to take the step from the exceptional case of perfect algorithms to imperfect algorithms. *Imperfect algorithms* make some assignments that do not match the ground truth. There are two kinds of reasons for this. First, classical reasons as over-/under-fitting, biased/lacking training data, missing features, etc. (Bishop, 2006). The second kind of reason is more principled than the first. Supervised learning algorithms lack the ability to distinguish between causes and correlations (Pearl and Mackenzie, 2018). Therefore, variables that only correlate but have no causal relationship with the target variable do, in fact, impact the classification. While there are various ways to get rid of the first kind of problems (Claeskens et al., 2008; Good and Hardin, 2012; Jabbar and Khan, 2015), getting causality into supervised learning is much harder (Schölkopf, 2019). Both kinds of reasons lead to classifications that mismatch the ground truth. Thus, in both cases, some features have an undesired impact on the target variable. This mismatch can be used in creating good AEs but also good CEs. The following example will show a scenario where only contesting CEs exist but no feasible CEs.

Scenario. For the sake of argument, assume that a bank collected data from the members of two clubs. The first club is a dog-club in Zurich (Switzerland) and the second is an animal protection club in Ukraine. It is clear that this data collection is biased. Let us also assume that the model trained by the bank is a single-layer decision tree. The algorithm has learned that the number of dogs is the only important feature for deciding on a loan application. If a person has more or equal to two dogs, the algorithm offers the loan.
Assume the loan applicant has a low salary s and one dog. In this case, the loan application would be rejected. This decision would be correct according to the ground truth since the salary was too low. However, the algorithm is right for the wrong reasons. It rejects the application because the applicant does not have at least two dogs. A CE, in this case, would be:

- If P's had one more dog, her loan application would have been accepted.

[34] This again points out that the class of CEs is broader than of AEs (See Appendix A + B).

This explanation would indeed be a good CE since it points us to the reason the algorithm had for its decision. It would increase the applicant's understanding of the algorithm and would allow her to contest the decision. Besides, in case she really urges money, she could use this information to deceive the algorithm. These functions exactly characterize contesting CEs. Interestingly, an AE would be described by the same vector $(s, 2)$ and can even have the same function - deceiving the system.

4.3. Mixed CEs

Usually, we neither deal with the perfect algorithms from Section 4.1 nor the terrible algorithms of Section 4.2. Instead, we often deal with algorithms that mostly focus on relevant features to the right amount but sometimes make misclassifications. In such scenarios, we can have feasible, contesting, and also mixed CEs.

Scenario. Consider again an imperfect algorithm, as discussed in Section 4.2. However, this time, the model selection and the data collection were carried out more carefully, and all potential kind one fallacies have been avoided. As a result, the algorithm has learned that the salary is relevant for loan acceptance. Also, dogs are expensive, and therefore only people with a comparatively high salary can afford dogs. Since the algorithm only matches patterns and cannot tell non-causal dependencies from actual causes apart, it will learn that the number of dogs is (slightly) relevant for loan acceptance.
Now, consider an applicant with a salary s very close to reaching the decision boundary t of loan acceptance and zero dogs. In accordance with the ground truth, the bank rejects the loan application since s is below t. However, the algorithm is not perfect. It learned that additionally to the salary, the number of dogs is marginally relevant for the applicant obtaining the loan. There is potentially a variety of CEs. Three possible CEs could be

i) If P's salary was $(t - s)$€ higher, her loan application would have been accepted.

ii) If P's had two more dogs, her loan application would have been accepted.

iii) If P's salary was $\frac{(t-s)}{2}$€ higher and she had one more dog, her loan application would have been accepted.

All of them would be good CEs. All of them would provide information for a better understanding of the algorithm. i) would be a feasible CE and point to the most relevant feature that is also causally relevant. ii) is a contesting CE. It points to a causally irrelevant feature that is important according to the algorithm. Moreover, it is the same vector as one possible good AE. iii) is a mixed type CE. It gives information about the most important feature but at the same time also about a secondary feature that should not matter but does. Similar to contesting CEs, it allows contestability but potentially also feasibility. It would also be an AE. However, not a good one because even though it is misclassified, it shows perceptible changes in the salary feature.

4.4. AEs as Contesting CEs

These examples raise the question of whether every good AE makes a good contesting CE and vice versa. Indeed, the two classes have a significant overlap. First, they both share the potential function of deception. Second, both provide grounds to contest the judgment of a machine learning algorithm. One difference is that as all CEs, contesting CEs highlight the changes. AEs, on the other hand, try to hide the changes as well as possible.
Every contesting CE is an AE as it must be misclassified to contest the decision for justified reasons. If there is a misclassification, there will potentially be contexts in which an attacker can exploit this bug. Hence, the most interesting contesting CEs will also be good AEs.

What about the opposite direction? There might be cases where a vector is a good AE but a not so good contesting CE. This case occurs when many causally irrelevant features are changed to achieve an alternative classification. However, it is unclear whether sparsity is a mandatory prerequisite for a good CE. In particular, if the agent aims to deceive a system via a contesting CE, she might not care too much about sparsity. Also, distributed changes in AEs are mainly relevant in the context of computer vision. However, as we have mentioned, for CEs, the interpretability of features is essential. In the case of images, one could argue that a minor change in all features means a change in only one interpretable feature, namely the image's coloration. The change of coloration as the only feature is, in fact, sparse and it can rightly be argued that it is not causally relevant for classification. Thus, contesting CEs and AEs have at least a considerable overlap, and it is difficult to find convincing cases that fit in one class but not in the other.

4.5. Causality as a Unifying Perspective

Many recently proposed papers on CEs have focused on feasibility and actionability for generating counterfactuals. They mostly achieved this aim by incorporating causal domain knowledge (Poyiadzi et al., 2020; Mahajan et al., 2019; Karimi et al., 2020). Since explanations should guide our future actions, this indeed makes sense. However, if an algorithm uses questionable features in its decisions or misuses features, we may want explanations that faithfully reflect such flaws. In such cases, we are interested in contesting CEs (that reassemble AEs), which point to causally irrelevant features that have influenced the learning algorithm's decision.

Summarized, we can say that distance measures to build good contesting CEs (AEs) assign small values to changes in impactful but causally irrelevant features. This setting implies that only causally irrelevant features are changed since we minimize the distance to the alternative input. Adequate distance measures for feasible CEs, on the other hand, assign low values to causally relevant and actionable features, whereby these features are altered strongest. Hence, we can say that there are at least two classes of interesting CEs that can, in some cases, be mixed. Class one are feasible CEs. They guide the explainee's future actions by the given recommendations and stand in accordance with the real-world causal structure. Class two are contesting CEs. They allow us to contest the decisions of algorithms or deceive them, just like AEs.

Can this idea of the two complementary approaches be transferred from tabular data to images? We believe that this is possible. The only difference is that the features we find causally relevant for the classification are composed of the input features the algorithm receives, namely pixels. Changing all pixels a little bit or a few pixels strongly is potentially correlated to a different classification; however, it is causally irrelevant. This idea relates to the paper of Ilyas et al. (2019), where they discuss predictive but non-robust features. They argue that these features are the reason for the occurrence of AEs. We think that one important subclass of such features are correlated, non-causal features.

5. Discussion

CEs and AEs are strongly related approaches. Our conceptual comparison has shown that their commonalities go deeper than the mathematical similarity alone. Did our analysis shed light on the three misconceptions discussed in Section 2?

The first misconception was to consider CE and AE as synonyms. Our analysis has shown that every (targeted) AE is a (targeted) CE, but not vice versa.[35]. The essential difference is that AEs, by definition, must

[35] For an example of a CE that is not an AE, see Appendix B

be misclassified, whereas CEs are in this respect agnostic. The second misconception was to consider feasible CEs as the only relevant type of CEs. We showed that contesting CEs are another important type of CEs different from feasible CEs. The difference in function between the two types appears in a difference in their notion of similarity between inputs. Contesting CEs target misclassified inputs and show therefore remarkable similarity with AEs. The third misconception concerned unjustified or missing transfers between the fields. We discussed under which conditions fruitful interactions are possible. While transfers of the optimization problem or the solution methods are mostly permissible, transfers on the respective distance notions are more demanding. Specifically, we argued that feasibility and misclassification are contrary aims, whereas informativeness and imperceptibility go well together.

6. Outlook

In addition to clarifying some misconceptions, this paper opens various directions for future research. First and foremost, it directs to a preference-based selection of feasible and contesting CEs, including degrees of feasibility/contestability. Second, as suggested, many concepts from AEs can be transferred and used in generating CEs. Especially if the domains show greater overlap, e.g., AEs for tabular data and CEs for image/audio data, such transfers will be beneficial. Conceptually the relation between the paradigm of supervised learning and the transferability of AEs needs further research.

Acknowledgements

Funding: This work was supported by the Graduate School of Systemic Neuroscience (GSN) of the LMU Munich. Big thanks to Stephan Hartmann, Christoph Molnar, Gunnar König, and the GSN Neurophilosophy-group for their helpful comments to the manuscript, the fruitful discussions about the concepts, and their hints to related literature.

References

Alzantot, M., Sharma, Y., Chakraborty, S., Zhang, H., Hsieh, C.J., Srivastava, M.B., 2019. Genattack: Practical black-box attacks with gradient-free optimization, in: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 1111–1119.

Asher, N., Paul, S., Russell, C., 2020. Adequate and fair explanations. arXiv preprint arXiv:2001.07578 .

Athalye, A., Engstrom, L., Ilyas, A., Kwok, K., 2018. Synthesizing robust adversarial examples, in: International conference on machine learning, PMLR. pp. 284–293.

Bach, F., 2010. Sparse methods for machine learning, in: Tutorial of IEEE-CS Conference on Computer Vision and Pattern Recognition (CVPR).

Balda, E.R., Behboodi, A., Mathar, R., 2019. Perturbation analysis of learning algorithms: generation of adversarial examples from classification to regression. IEEE Transactions on Signal Processing 67, 6078–6091.

Ballet, V., Renard, X., Aigrain, J., Laugel, T., Frossard, P., Detyniecki, M., 2019. Imperceptible adversarial attacks on tabular data. arXiv preprint arXiv:1911.03274 .

Barocas, S., Selbst, A.D., Raghavan, M., 2020. The hidden assumptions behind counterfactual explanations and principal reasons, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA. p. 80–89. URL: https://doi.org/10.1145/3351095.3372830, doi:10.1145/3351095.3372830.

Beaney, M., 2018. Analysis, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. summer 2018 ed.. Metaphysics Research Lab, Stanford University.

Bekoulis, G., Deleu, J., Demeester, T., Develder, C., 2018. Adversarial training for multi-context joint entity and relation extraction. arXiv preprint arXiv:1808.06876 .

Bishop, C.M., 2006. Pattern recognition and machine learning. Springer.

Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J., 2017. Adversarial patch. arXiv preprint arXiv:1712.09665 .

Carlini, N., Katz, G., Barrett, C., Dill, D.L., 2018. Ground-truth adversarial examples. URL: https://openreview.net/forum?id=Hki-ZlbA-.

Carlini, N., Wagner, D., 2017. Towards evaluating the robustness of neural networks, in: 2017 IEEE Symposium on Security and Privacy, IEEE. pp. 39–57.

Carnap, R., 1998. Der logische Aufbau der Welt. volume 514. Felix Meiner Verlag.

Chabris, C.F., Simons, D.J., 2010. The invisible gorilla: And other ways our intuitions deceive us. Harmony.

Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J., 2017. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models, in: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, pp. 15–26.

Claeskens, G., Hjort, N.L., et al., 2008. Model selection and model averaging. Cambridge Books doi:10.1017/CBO9780511790485.

Dalvi, N., Domingos, P., Sanghai, S., Verma, D., 2004. Adversarial classification, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 99–108.

Dandl, S., Molnar, C., Binder, M., Bischl, B., 2020. Multi-objective counterfactual explanations. arXiv preprint arXiv:2004.11165 .

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 .

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., Sohl-Dickstein, J., 2018. Adversarial examples that fool both computer vision and time-limited humans, in: Advances in Neural Information Processing Systems, pp. 3910–3920.

Fernández-Loría, C., Provost, F., Han, X., 2020. Explaining data-driven decisions made by ai systems: The counterfactual approach. arXiv:2001.07417.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2015. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24, 44–65. doi:10.1080/10618600.2014.907095.

Good, P.I., Hardin, J.W., 2012. Common errors in statistics (and how to avoid them). John Wiley & Sons. doi:10.1002/9781118360125.

Goodfellow, I., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples, in: International Conference on Learning Representations. URL: http://arxiv.org/abs/1412.6572.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F., 2018. Local rule-based explanations of black box decision systems. arXiv preprint arXiv:1805.10820 .

Guo, C., Gardner, J.R., You, Y., Wilson, A.G., Weinberger, K.Q., 2019. Simple black-box adversarial attacks. arXiv preprint arXiv:1905.07121 .

Ignatiev, A., Narodytska, N., Marques-Silva, J., 2019. On relating explanations and adversarial examples, in: Advances in Neural Information Processing Systems, pp. 15883–15893.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., Madry, A., 2019. Adversarial examples are not bugs, they are features, in: Advances in Neural Information Processing Systems, pp. 125–136.

Ioannou, C.I., Pereda, E., Lindsen, J.P., Bhattacharya, J., 2015. Electrical brain responses to an auditory illusion and the impact of musical expertise. PLoS One 10.

Jabbar, H., Khan, R.Z., 2015. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Computer Science, Communication and Instrumentation Devices doi:10.3850/978-981-09-5247-1_017.

Jehee, J.F., Brady, D.K., Tong, F., 2011. Attention improves encoding of task-relevant features in the human visual cortex. Journal of Neuroscience 31, 8210–8219.

Kahneman, D., Slovic, S.P., Slovic, P., Tversky, A., 1982. Judgment under uncertainty: Heuristics and biases. Cambridge University Press.

Kanamori, K., Takagi, T., Kobayashi, K., Arimura, H., 2020. Dace: Distribution-aware counterfactual explanation by mixed-integer linear optimization, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, pp. 2855–2862.

Karimi, A.H., Schölkopf, B., Valera, I., 2020. Algorithmic recourse: from counterfactual explanations to interventions, in: 37th International Conference on Machine Learning (ICML).

Kusner, M.J., Loftus, J., Russell, C., Silva, R., 2017. Counterfactual fairness, in: Advances in Neural Information Processing Systems, pp. 4066–4076.

Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M., 2019. The dangers of post-hoc interpretability: Unjustified counterfactual explanations, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization. pp. 2801–2807. URL: https://doi.org/10.24963/ijcai.2019/388, doi:10.24963/ijcai.2019/388.

Lewis, D.K., 1973. Counterfactuals. Blackwell.

Lu, J., Issaranon, T., Forsyth, D., 2017. Safetynet: Detecting and rejecting adversarial examples robustly, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 446–454.

Mahajan, D., Tan, C., Sharma, A., 2019. Preserving causal constraints in counterfactual explanations for machine learning classifiers. arXiv preprint arXiv:1912.03277 .

Menzies, P., Beebee, H., 2019. Counterfactual theories of causation, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. winter 2019 ed.. Metaphysics Research Lab, Stanford University.

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

Molnar, C., 2019. Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B., 2020. Pitfalls to avoid when interpreting machine learning models. arXiv:2007.04131.

Moore, J., Hammerla, N., Watkins, C., 2019. Explaining deep learning models with constrained adversarial examples, in: Pacific Rim International Conference on Artificial Intelligence, Springer. pp. 43–56.

Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P., 2017. Universal adversarial perturbations, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1765–1773.

Mothilal, R.K., Sharma, A., Tan, C., 2020. Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency.

Navalpakkam, V., Itti, L., 2005. Modeling the influence of task on attention. Vision research 45, 205–231.

Páez, A., 2019. The pragmatic turn in explainable artificial intelligence (xai). Minds and Machines 29, 441–459.

Papernot, N., McDaniel, P., Goodfellow, I., 2016a. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:1605.07277 .

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A., 2017. Practical black-box attacks against machine learning, in: Proceedings of the 2017 ACM on Asia conference on computer and communications security, pp. 506–519.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A., 2016b. The limitations of deep learning in adversarial settings, in: 2016 IEEE European symposium on security and privacy (EuroS&P), IEEE. pp. 372–387.

Pearl, J., 2009. Causality. Cambridge University Press.

Pearl, J., Mackenzie, D., 2018. The book of why: the new science of cause and effect. Basic Books.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., Flach, P., 2020. Face: Feasible and actionable counterfactual explanations, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 344–350.

Radford, A., Metz, L., Chintala, S., 2016. Unsupervised representation learning with deep convolutional generative adversarial networks, in: Bengio, Y., LeCun, Y. (Eds.), 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings. URL: http://arxiv.org/abs/1511.06434.

Reutlinger, A., 2018. Extending the counterfactual theory of explanation. pp. 74–95. doi:10.1093/oso/9780198777946.003.0005.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. 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, ACM. pp. 1135–1144. doi:10.1145/2939672.2939778.

Rozsa, A., Rudd, E.M., Boult, T.E., 2016. Adversarial diversity and hard positive generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 25–32.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1, 206–215.

Russell, B., 1905. On denoting. Mind 14, 479–493.

Russell, C., 2019. Efficient search for diverse coherent explanations, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA. p. 20–28. URL: https://doi.org/10.1145/3287560.3287569, doi:10.1145/3287560.3287569.

Sabour, S., Cao, Y., Faghri, F., Fleet, D.J., 2016. Adversarial manipulation of deep representations, in: Bengio, Y., LeCun, Y. (Eds.), 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings. URL: http://arxiv.org/abs/1511.05122.

Schölkopf, B., 2019. Causality for machine learning. arXiv preprint arXiv:1911.10500 .

Sharma, S., Henderson, J., Ghosh, J., 2020. Certifai:: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society URL: http://dx.doi.org/10.1145/3375627.3375812, doi:10.1145/3375627.3375812.

Sokol, K., Flach, P.A., 2019. Counterfactual explanations of machine learning predictions: Opportunities and challenges for ai safety, in: Proceedings of the AAAI Workshop on Artificial Intelligence Safety.

Starr, W., 2019. Counterfactuals, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. fall 2019 ed.. Metaphysics Research Lab, Stanford University.

Štrumbelj, E., Kononenko, I., 2014. Explaining prediction models and individual predictions with feature contributions. Knowledge and information systems 41, 647–665. doi:10.1007/s10115-013-0679-x.

Stutz, D., Hein, M., Schiele, B., 2019. Confidence-calibrated adversarial training: Generalizing to unseen attacks. arXiv preprint arXiv:1910.06259 .

Su, J., Vargas, D.V., Sakurai, K., 2019. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation 23, 828–841.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R., 2014. Intriguing properties of neural networks, in: International Conference on Learning Representations. URL: http://arxiv.org/abs/1312.6199.

Tomsett, R., Widdicombe, A., Xing, T., Chakraborty, S., Julier, S., Gurram, P., Rao, R., Srivastava, M., 2018. Why the failure? how adversarial examples can provide insights for interpretable machine learning, in: 21st International Conference on Information Fusion (FUSION), IEEE. pp. 838–845.

Ustun, B., Spangher, A., Liu, Y., 2019. Actionable recourse in linear classification, in: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 10–19.

Van Looveren, A., Klaise, J., 2019. Interpretable counterfactual explanations guided by prototypes. arXiv preprint arXiv:1907.02584 .

Wachter, S., Mittelstadt, B., Russell, C., 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. 31, 841.

Wheeler, G., 2020. Bounded rationality, in: Zalta, E.N. (Ed.), The Stanford Encyclopedia of Philosophy. spring 2020 ed.. Metaphysics Research Lab, Stanford University.

Yuan, X., He, P., Zhu, Q., Li, X., 2019. Adversarial examples: Attacks and defenses for deep learning. IEEE transactions on neural networks and learning systems 30, 2805–2824.

Appendix A. Formal Proof for *AEs* ⊆ *CEs*

In this section, we consider the relation between CEs and AEs in purely mathematical terms. For all the following, assume there is a function $f : I \to O$ mapping a vector x from a (potentially high dimensional) input space I to a vector y in an output space O.

Definition. Let $x \in I$, $f(x) = y \in O$ and $y' \in O$.

- We call $x'_x \in I$ an *alternative* to x if $y \neq f(x')$.
- We call $x'_{x,y'} \in I$ a *targeted alternative* to x if $f(x'_{x,y'}) = y' \neq y$.

Definition. A distance measure $\|\cdot\|$ on a space I is defined as a function $\|\cdot\| : I \to \mathbb{R}_+ \cup \{\infty\}$.

Definition. Let $x \in I$ be a vector, $x'_x \in I$ be an alternative vector, $\epsilon > 0$, and $\|\cdot\|$ be a distance measure on I.

- We call $x'_{x,\epsilon}$ an *ϵ-alternative* to x with respect to $\|\cdot\|$ if $\|x - x'_x\| < \epsilon$.
- We call $x'_{x,y',\epsilon}$ a *targeted-ϵ-alternative* to x with respect to $\|\cdot\|$ and target class y' if $x'_{x,y',\epsilon}$ is a targeted-alternative and an ϵ-alternative.

Definition. Let $x \in I$, $f(x) \neq y' \in O$, $\epsilon > 0$, and $\|\cdot\|$ be a distance measure on I.

- Let $A(x)$ be the set of all alternatives to x.
- Let $TA(x, y')$ be the set of all targeted-alternatives to x with target class y'.
- Let $A(x, \epsilon, \|\cdot\|)$ be the set of all ϵ-alternatives to x with respect to $\|\cdot\|$.
- Let $TA(x, y', \epsilon, \|\cdot\|)$ be the set of all targeted ϵ-alternatives to x with respect to $\|\cdot\|$ and the target class y'.

Theorem. *For all $x \in I$, $\epsilon > 0$, $f(x) \neq y' \in O$ and distance measures $\|\cdot\|$ on I holds:*

i) $A(x, \epsilon, \|\cdot\|) \subseteq A(x)$

ii) $TA(x, y', \epsilon, \|\cdot\|) \subseteq TA(x, y')$

iii) $TA(x, y', \epsilon, \|\cdot\|) \subseteq A(x, \epsilon, \|\cdot\|)$

iv) $TA(x, y') \subseteq A(x)$

v) $\forall \delta > \epsilon : A(x, \epsilon, \|\cdot\|) \subseteq A(x, \delta, \|\cdot\|)$

vi) $\forall \delta > \epsilon : TA(x, y', \epsilon, \|\cdot\|) \subseteq TA(x, y', \delta, \|\cdot\|)$

Proof.

i) Let $x \in I$, $\epsilon > 0$, $\|\cdot\|$ and $z \in A(x, \epsilon, \|\cdot\|)$ be arbitrary. Then, $f(z) \neq f(x)$. Thus, $z \in A(x)$.

ii) Let $x \in I$, $\epsilon > 0$, $f(x) \neq y' \in O$, $\|\cdot\|$ and $z \in TA(x, y', \epsilon, \|\cdot\|)$ be arbitrary. Then, $f(z) = y' \neq f(x)$. Thus, $z \in TA(x, y')$.

iii) Let $x \in I$, $\epsilon > 0$, $f(x) \neq y' \in O$, $\|\cdot\|$ and $z \in TA(x, y', \epsilon, \|\cdot\|)$ be arbitrary. Then, $f(z) = y' \neq f(x)$ and $\|x - z\| < \epsilon$. Thus, $z \in A(x, \epsilon, \|\cdot\|)$.

iv) Let $x \in I$, $f(x) \neq y' \in O$, and $z \in TA(x, y')$ be arbitrary. Then, $f(z) = y' \neq f(x)$. Thus, $z \in A(x)$.

v) Let $x \in I$, $\epsilon < \delta$, $\|\cdot\|$ be arbitrary and $z \in A(x, \epsilon, \|\cdot\|)$. Then, by definition $\|x - z\| < \epsilon$ and $f(z) \neq f(x)$. Together with transitivity in the real numbers follows $\|x - z\| < \delta$. Thus, $z \in A(x, \delta, \|\cdot\|)$.

vi) Let $x \in I$, $f(x) \neq y' \in O$, $\epsilon < \delta$, $\|\cdot\|$ be arbitrary and $z \in TA(x, y', \epsilon, \|\cdot\|)$. Then, by definition $\|x - z\| < \epsilon$ and $f(z) = y' \neq f(x)$. Together with transitivity in the real numbers follows $\|x - z\| < \delta$. Thus, $z \in TA(x, y', \delta, \|\cdot\|)$.

□

Definition. Let $\epsilon > 0$

- We call x_c a *non-targeted ϵ counterfactual* to x with respect to $\|\cdot\|$ if $x_c \in A(x, \epsilon, \|\cdot\|)$. We call x_c a *non-targeted counterfactual* if it is a non-targeted ϵ counterfactual and for all $\delta < \epsilon$ holds $A(x, \delta, \|\cdot\|) = \emptyset$.
- We call $x_{c,y'}$ a *targeted ϵ counterfactual* to x with respect to $\|\cdot\|$ and targetclass y' if $x_{c,y'} \in TA(x, y', \epsilon, \|\cdot\|)$. We call $x_{c,y'}$ a *targeted counterfactual* if it is a targeted ϵ counterfactual and and for all $\delta < \epsilon$ holds $TA(x, y', \delta, \|\cdot\|) = \emptyset$.

Definition. Let $x \in I$, $f(x) \neq y' \in O$ and $\|\cdot\|$ be a distance measure:

- We call x_a a *non-targeted (ϵ) adversarial example* to x with respect to $\|\cdot\|$ if x_a is a non-targeted (ϵ) counterfactual and there exists a ground truth $y_{GT} \in O$ for x_a such that $f(x_a) \neq y_{GT}$.
- We call $x_{a,y'}$ a *targeted (ϵ) adversarial example* to x with respect to $\|\cdot\|$ and target class y' if $x_{a,y'}$ is a targeted (ϵ) counterfactual and there exists a ground truth $y_{GT} \in O$ for $x_{a,y'}$ such that $y' = f(x_a) \neq y_{GT}$.

Definition. Let $x \in I$, $f(x) \neq y' \in O$, $\epsilon > 0$, and $\|\cdot\|$ be a distance measure on I.

- Let $CE(x, \epsilon, \|\cdot\|)$ be the set of all non-targeted ϵ counterfactuals to x with respect to $\|\cdot\|$. Let $CE(x, \|\cdot\|)$ be the set of all non-targeted counterfactuals to x with respect to $\|\cdot\|$.
- Let $TCE(x, y', \epsilon, \|\cdot\|)$ be the set of all targeted ϵ counterfactuals to x with respect to $\|\cdot\|$ and target class y'. Let $TCE(x, y', \|\cdot\|)$ be the set of all targeted counterfactuals to x with respect to $\|\cdot\|$ and target class y'.
- Let $AE(x, \epsilon, \|\cdot\|)$ be the set of all non-targeted ϵ adversarial examples to x with respect to $\|\cdot\|$. Let $AE(x, \|\cdot\|)$ be the set of all non-targeted adversarial examples to x with respect to $\|\cdot\|$.
- Let $TAE(x, y', \epsilon, \|\cdot\|)$ be the set of all targeted ϵ adversarial examples to x with respect to $\|\cdot\|$ and target class y'. Let $TAE(x, y', \|\cdot\|)$ be the set of all targeted adversarial examples to x with respect to $\|\cdot\|$ and target class y'.

Theorem (Every (targeted ϵ) AE is a (targeted ϵ) CE). *For all $x \in I$, $f(x) \neq y' \in O$, $\epsilon > 0$ and distance measures $\|\cdot\|$ holds:*

i) $AE(x, \epsilon, \|\cdot\|) \subseteq CE(x, \epsilon, \|\cdot\|)$

ii) $TAE(x, y', \epsilon, \|\cdot\|) \subseteq TCE(x, y', \epsilon, \|\cdot\|)$

iii) $AE(x, \|\cdot\|) \subseteq CE(x, \|\cdot\|)$

iv) $TAE(x, y', \|\cdot\|) \subseteq TCE(x, y', \|\cdot\|)$

Proof. All statements follow directly by the definition of adversarials given above. □

This Theorem is not true if we consider an ϵ-environment for (non-)targeted CEs and an δ-environment for (non-)targeted AEs where $\epsilon \neq \delta$.

Appendix B. Example CE but not AE

Figure B.4 from Van Looveren and Klaise (2019) shows an example of image data CEs that are not AEs. The nine we see is a counterfactual to the original input. It is a variation of the original input that points us to the crucial difference between an eight and a nine, which is the lower-left stroke. Now, we look at it from an adversarial perspective. The eight looks like an eight. The alternative generated to the eight, which is classified as a nine, does look like a nine. This example cannot be an AE because it is not a misclassification.

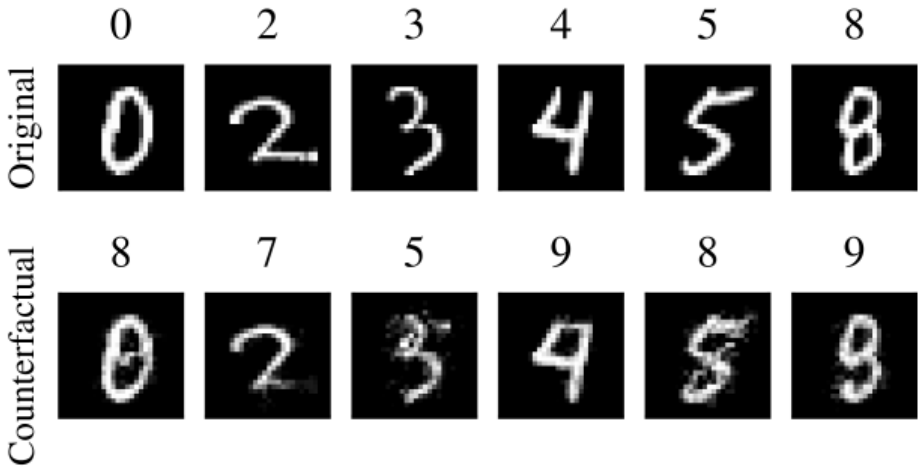

Figure B.4: Above the two pictures, you can see the corresponding classes the algorithm has assigned. The picture classified as an eight is the original data input. The nine beneath is a counterfactual to the eight.