Counterfactual Explanations & Adversarial Examples

Common Grounds, Essential Differences, and Potential Transfers

Timo Freiesleben

Munich Center for Mathematical Philosophy &
Graduate School of Systemic Neurosciences, LMU Munich.

Abstract

It is well known that adversarial examples and counterfactual explanations are based on the same mathematical model. However, their relationship has not yet been studied at a conceptual level. The present paper fills this gap. We show that counterfactual reasoning is the common basis of the fields and reliable machine learning their shared goal. Moreover, we illustrate to what extent counterfactual explanations can be regarded as the more general concept than adversarial examples. We introduce the conceptual distinction between feasible and contesting counterfactual explanations and argue that adversarial examples are similar to the latter.

Keywords: Counterfactual Explanation, Adversarial Example, XAI, Causality

Contents

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]. The solution to this problem has evolved into an area called eXplainable Artificial Intelligence (XAI) or Interpretable Machine Learning (IML). A variety of interpretation techniques have been proposed, among which model agnostic methods e.g. ICE curves [Goldstein et al., 2015], LIME [Ribeiro et al., 2016], Shapley values [Štrumbelj and Kononenko, 2014], etc. have recently gained popularity as they do not pose any assumptions on the employed model to be applicable [Molnar, 2019]. Counterfactual Explanation (CE) [Wachter et al., 2017] is one of these model agnostic methods. It aims to explain particular 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]. Especially in the field of computer vision, it has been shown that successful attacks can be constructed for almost any learning algorithm [Goodfellow et al., 2015, 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, they 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. [2019] use counterfactuals in their measure of robustness 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 research.

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 contexts of use. In order to compare CEs and AEs, we need to analyze each of the two fields. For this analyses, it is important to focus on aspects that are sufficient to describe both fields [Beaney, 2018].[2] Moreover, the aspects should allow to lead an informed discussion about

---

[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

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 and contexts in which the concepts are used. 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.

We start our discussion in Section 2 with three potential misconceptions researchers on CEs and AEs should be aware of. 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. It is argued 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 serious confusions. The main goal of this analysis is therefore to resolve these confusions and lay the foundation for well-guided future research on both CEs and AEs.

CE is equal to AE: Misconception number one can be summarized as follows: *As CEs and AEs share the same mathematical model, they are the same object.* Authors with this idea in mind like Sharma et al. [2019] use the terms counterfactual and adversarial interchangeably. This is a misconception regarding the very basis of CEs and AEs. Not only do CEs and AEs have partially non-overlapping goals, but also must AEs satisfy an additional constraint, which CEs do not have to satisfy. This constraint is the misclassification of the adversarial. This misconception may lead to a false interpretation of robustness against attacks and thereby to worse performance and societal acceptance of machine learning applications. We examine misclassification in detail in Section 3.1 and Section 3.3.

---

the concepts, their use cases, state of the art, etc.

4

Feasible CEs are all CEs: A second misconception appears in the context of CEs. It can be summarized as follows: *We want CEs that make sense and that guide our future actions. Hence, these are the types of CEs that should be generated for end-users.* Researchers focusing on feasibility/actionability of CEs [Poyiadzi et al., 2020, Mahajan et al., 2019, Karimi et al., 2020] do not usually claim that these are the only relevant types of CEs. However, they neither discuss the problems of focusing on such CEs nor discuss potential other types of CEs. One type that is thereby left out are CEs that focus on the contestability of a decision. Such CEs provide ground for end-users to contest a decision and have therefore major legal relevance. Interestingly, contesting CEs show a great resemblance with AEs. If one were to focus only on feasible CEs, this would lead to hiding biased algorithmic decisions behind the facade of an explanation. A discussion of feasibility and contestability is given in Section 3.2 and the distinction between feasible and contesting CEs is introduced in Section 4.

Transfers, yes or no?: The third misconception is twofold. It either goes like this: *CEs and AEs seem mathematically very similar. So models and algorithms developed for one side can be fruitfully applied to the other side as well.* or like this: *CEs and AEs have nothing to do with each other. Therefore, researchers on CEs/AEs can ignore the work on AEs/CEs.* In the former case, methods can be misused or hidden assumptions can be adopted..[3] In the latter case, already known techniques are potentially rediscovered.[4] To avoid these transfer problems, we will discuss about the conceptually permissible transfers between CEs and AEs in Section 3.3 in detail.
We will come back to these three misconceptions in Section 5.

3 CEs & AEs: A Comparison

To give the reader an intuition, 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 online portal of a bank. 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.

---

[3]E.g Generating counterfactuals based on AE surrogate techniques [Guidotti et al., 2018] uses local approximations and faces therefore the same critiques as LIME [Molnar, 2019].

[4]E.g. evolutionary algorithms for mixed data CEs [Sharma et al., 2019].

AE example: Look at Figure 1 [Papernot et al., 2017]. What makes the modified images in the second-row adversarial examples is that the small modifications changed the classification, while for humans the images look almost the same. Therefore, this change in prediction is unjustified and the system 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. In the row below, we see the same five pictures but slightly modified by some noise added to the pictures. Here, they are misclassified.

3.1 Basis

The first aspect we want to investigate is the conceptual basis of CEs and AEs. We first consider each of them in separation and then conduct our comparison.

Counterfactual Explanations

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

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

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 generally accepted solution has been found [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

---

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

$\omega \in \Omega$ 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 counterfactual explanation that is helpful to the person for whom the explanation is generated. This means that the counterfactual explanation is easy to comprehend and has an interesting[8] antecedent/consequent for the explainee [Miller, 2019]. Usually, in XAI application contexts, 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]. This is mostly not the goal of these types 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 about 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 it 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. These meaningless inputs are not considered AEs. What defines adversarial examples is that the inputs we give 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 are based on a classical picture from game theory, in which a deceiver-agent tries to trick a discriminator-agent [Dalvi et al., 2004].[11] An AE denotes a case where the deceiver was successful. Adversarial attacks are not specific to neural networks [Papernot et al., 2016a]. Almost any complex enough system can be tricked, and even humans are not invulnerable to that [Kahneman et al., 1982, Chabris and Simons, 2010, Ioannou et al., 2015]. What is special about modern AEs is that they can often be transferred from one model to another and that it is particularly hard to get rid of them [Yuan et al., 2019]. The origin of this effect is still open to debate [Goodfellow et al., 2015, Ilyas et al., 2019].

---

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

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

[8]This intuitive notion of interestingness is specified in Section 3.2.

[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].

Comparison

Both fields rely on counterfactual reasoning. CEs describe a variation of the actual situation/world. AEs are a variation of a real-world input. Also, in both cases, the variation changes the result. Furthermore, both approaches search for this variation in regions close to the real world (input) which fulfill 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 would have to satisfy that a predicate Q applies. For AEs, an alternative to a real-world input must be misclassified compared to some ground truth. We can already see how the later is a specific case of the former where Q is defined as the predicate 'being misclassified'. Misclassification is generally not demanded for CEs. Conversely, CEs often demand that Q describes a specific outcome as in our example the '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.

Not all three of these goals need to be met together for an explanation to be considered good. In many contexts, explanations focus on only one or two of the three goals.

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, central reasons for a decision. To achieve this degree of simplicity in our explanations, we must be economical with respect to the number of reasons we give. For this reason, *sparsity* is one aim when CEs are discussed in the literature.

---

[12]By "decisions" we usually mean classification or regression tasks the algorithm was optimized for.

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

Aim ii): Explanations should serve as a guideline for future actions. Hence, the alternative that reaches the desired output (e.g. obtaining the loan) should not be unreachable for the explainee. For example, a loan applicant cannot become younger in order to obtain a loan, even if age is one of the bank's criteria for justified reasons. Thus, it does not seem to be a reasonable recommendation to propose a reduction in age for obtaining a loan. Such limitations on our explanations are summarized in the literature under the term *feasibility*.

Aim iii): Explanations provide grounds for the appealability of decisions. A decision can be contested if the reasons given are poor. This may be because the decision is based on features that should not play a role (e.g. skin color or gender), or on features that should play a role but are expected to have a different effect (e.g. if a high salary is negatively correlated with obtaining a loan). All in all, we 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]

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 often 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 play an important role in XAI.

Use Cases: The use cases of CEs are in generally unlimited. In principle, CE can be applied to any kind of algorithm. The only requirements are that the input and output space is interpretable and that we can define a reasonable distance measure on these spaces. Despite that, CEs are almost exclusively considered 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 actually be applied on a broad scale is at least questioned [Laugel et al., 2019, Barocas et al., 2020].

---

[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].

[15]One counterexample to this is the work on the MNIST dataset by Van Looveren and Klaise [2019].

Adversarial Examples

The aim of AEs depends to a large extent on the perspective taken and the intended use case. This can be the perspective of the academic researcher working on AEs, the engineer trying to protect against attacks, or the attacker trying to abuse the system. The principal aims that all perspectives share are

i) to fool the system,

ii) to do this imperceptibly. and

iii) effectively.

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. What do the aims imply for our approach to the problem?

Aim i): The main aim of AEs is to fool the system. That is, we look for *missclassifications*. The system usually performs fairly well on training data and similar inputs. On the other hand, it performs poorly on inputs that are in unseen regions. If we look at any input 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. the changes in input should be below the threshold of human perception. This guarantees the highest chance of deceiving successfully. Since human perception directs attention to certain features and expectations, imperceptibility can be achieved by changes in features that are not payed attention to. Imperceptibility can also be achieved by keeping the number of changed features or the intensity of changes below the threshold of perception.

Aim iii): What makes an effective AE depends strongly on the context. Attackers want to exploit mistakes in the most profitable way possible (e.g. money 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, a depiction of real-world dangers in employing algorithms, and high research impact.

---

[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 great 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 AEs on them 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 our model [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 can help to understand what the algorithm has learned. Moreover, both contribute to the identification of biases and even offer methods to eliminate these biases through adversarial- or counterfactual-training [Bekoulis et al., 2018, Sharma et al., 2019]. 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 modifications in CEs that aim at informativeness. Imperceptible AEs result from modifications in unnoticed/unanticipated features which nevertheless influence the result. These surprisingly effective changes are exactly those which are most informative to humans, as will be discussed in Section 4.

In order for the feature permutations in CEs to make sense, it is usually necessary that the input space provides a certain degree of interpretability. This is not relevant for AEs, since the changes are hidden and not highlighted.[20]

---

[18]On case of AEs for tabular data give Ballet et al. [2019].

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

[20]One difference that might be pointed out is that CEs generally only tell us something about how an algorithm works in a very local region around a particular example while finding an AE affects confidence in the system as a whole. However, in cases where the CEs reveal racist, sexist, or causally unjustified reasons, this will also reduce our confidence in the system as a whole, not just locally.

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. In addition, both fields have gained increasing legal relevance. In order to be legally applicable (e.g. in autonomous driving, airport security [Athalye et al., 2017], etc.), machine learning algorithms must be both robust against AEs and provide explanations to end-users as specified in the GDPR.[21] A major difference is that AEs, by definition, can only point to mistakes of the algorithm. Hence, emerging AEs mainly have a negative role, while CEs can also raise trust in the system.

Considering the use cases, it can be seen that AEs are almost exclusively considered for computer vision tasks, while CEs consider almost exclusively tabular data as input.[22]

3.3   Models and Implementations

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

Counterfactual Explanations

There are a variety of formulations of the CE framework. The present version orients at Wachter et al. [2017]. Assume there is a *learning algorithm*[23], 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

$$\operatorname*{argmin}_{x' \in I} \|x - x'\| + \lambda \, \||f(x') - y'\|| \tag{2}$$

where $\| \cdot \|$ and $\|| \cdot \||$ are induced by some measures of distance on I and O respectively.[24] 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

---

[21] CEs comply with the standards given by the right to explanation in the GDPR [Wachter et al., 2017].

[22] Even though there are the above mentioned counterexamples to that division [Van Looveren and Klaise, 2019, Ballet et al., 2019].

[23] Usually this algorithm is already trained.

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

- 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 our input space. As discussed in Section 3.2, good CEs are sparse, feasible, and informative.

Sparsity is well studied in the general machine learning literature in different areas [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 a number of ways and depends on the perspective or problem you start with. One possibility is to declare some features as immutable, which makes 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 proposed as reasonable CE. In the literature, there are several suggestions how to deal with this problem, for example, by considering the probability density of inputs [Sharma et al., 2019], 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].[26]

While sparsity and feasibility are discussed throughout the literature, informativeness has so far not been taken into account. The reason might be that for informativeness we demand information about the explainee's estimates/expectations to which we usually do not have access. 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 prior characteristics by the average human and then to update this prior via feedback of the human agent.

Not only on the input space it might be difficult to find a suitable measure of distance but also on the output space. Consider a classification problem, where the output space is given by a set of probability density functions on the different categories. If the desired output is "loan application accepted" it is unclear whether this means that 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

---

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

[26]They combine blocking inpermutable features and avoiding unrealistic inputs.

loan is better than obtaining no loan. Standard measures like KL-divergence or cross-entropy are ignorant to such similarity differences.

Solution Methods: The solution strategy for the optimization problem depends on the model-knowledge. As the developers of IML techniques are usually also to 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]. Genetic algorithms are a solution method that does not require model knowledge [Sharma et al., 2019, Dandl et al., 2020]. A more controversial, technique that works 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].[27]

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 for instance 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. It is possible, for example, that loan applications from people below a certain salary level are subsidized by the state. 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]. But then the question arises, how many and which ones? Others propose to select a certain CE according to 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 in the case of CE, we will mainly focus on the deviations from it. 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. For AEs, no interpretability on the input space is required. There are two cases. Either, a

---

[27]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.

particular alternative output $y' \neq y$ is desired as in the case of CEs, which is called a *targeted attack*. Or, the alternative output just has to differ from y in which case we talk about a *non-targeted attack*. 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 that minimizes the term $\|x_{a,y'} - x\|$ for which $f(x_{a,y'}) = y'$. The formulation as a single objective optimization problem is again given by Equation (2). 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 inputs in greater distance might not only be computationally easier but also more interesting in some cases [Elsayed et al., 2018].

More important is that this input is in fact misclassified. Notice that this is guaranteed by none of the above formulations as optimization problems. To achieve this, we must add the condition that the alternative input is incorrectly classified.[28] 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. Clearly, this true label is usually the same as the one for our original input x, namely y.

The optimization problem presented here is only one among many others [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 important problems in creating an AE is the reasonable definition of a distance measure on the input space. We must be computationally able to minimize this measure, but, it should also allow us to find good AEs. This leads us to the aims of misclassification, imperceptibility, and effectiveness.

Again, minimizing the difference between x and its adversarial x_a, with the effect of flipping the algorithms assignment, does not guarantee to attain an AE. A switch in classification due to a small variation may be justified.[29] However, since we are often dealing with image data, a tiny variation[30] 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 you to an AE. Another perspective on the problem of misclassification can be found in Section 4.

---

[28]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].

[29]An example, is a case where a loan application for person P with the age of 17 years and 364 days is rejected while a person with the same characteristics but three days older would receive the loan.

[30]Often this variation is not only small in the sense of a p-norm, it is moreover structureless noise.

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. [2017]. 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., 2017]. This also leads to an overlap between human and machine deceivability [Elsayed et al., 2018]. With tabular data, it is much harder to define what imperceptibility is. Ballet et al. [2019] solved this via defining critical and non-critical features[31]. Since the algorithm uses both types of features in the classification, they modify only non-critical features to attain a change in assignment. This 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 main focus of the community is on the algorithmic generation of AEs, which again differs dependent on model knowledge. For white boxes, there are gradient-based methods, either for solving the optimization problem [Szegedy et al., 2014, Athalye et al., 2017, 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 a number of 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 are faced with two selection problems, neither of which has been discussed in the literature so far. Both depend on the desired effect. The *first selection problem* is: Based on which original input vector should the AE be generated? The *second selection problem* is: Given we get several AEs as solutions of the optimization problem, which AEs should we select?

A solution to the first selection problem relates to effectiveness and depends on three things: What is the application? Why do we want to deceive? Which are the weaknesses of the system? If we consider e.g. image recognition from a researcher's perspective and we want to illustrate the problems of autonomous driving, it makes a lot of sense to focus on safety issues. In cases like fraud, the

---

[31]Based on expert evaluation

choice is determined by the weaknesses of the system and its susceptibility to exploitation. From an engineering perspective, where protection is central, AEs are selected that pose the biggest threat.

The second selection problem we face is similar to the one in the case of CEs. Again, we should pick the AE that best suits our needs. However, since here the deceiver can pick the AE herself without constraints in resources this is not a very central problem.

Comparison

There is no need for researchers to reinvent the wheel. Since the two mathematical frameworks are very closely linked, we will additionally to our conceptual comparison, discuss potential transfers between the fields shown in 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 environment ϵ. However, it is important to point out here 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 in the distance measure. For that reason, we find the biggest 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 pointed at. Distance measures that favor sparsity can, therefore, be desirable to transfer between the fields. Moreover, distributed changes to achieve the imperceptibility
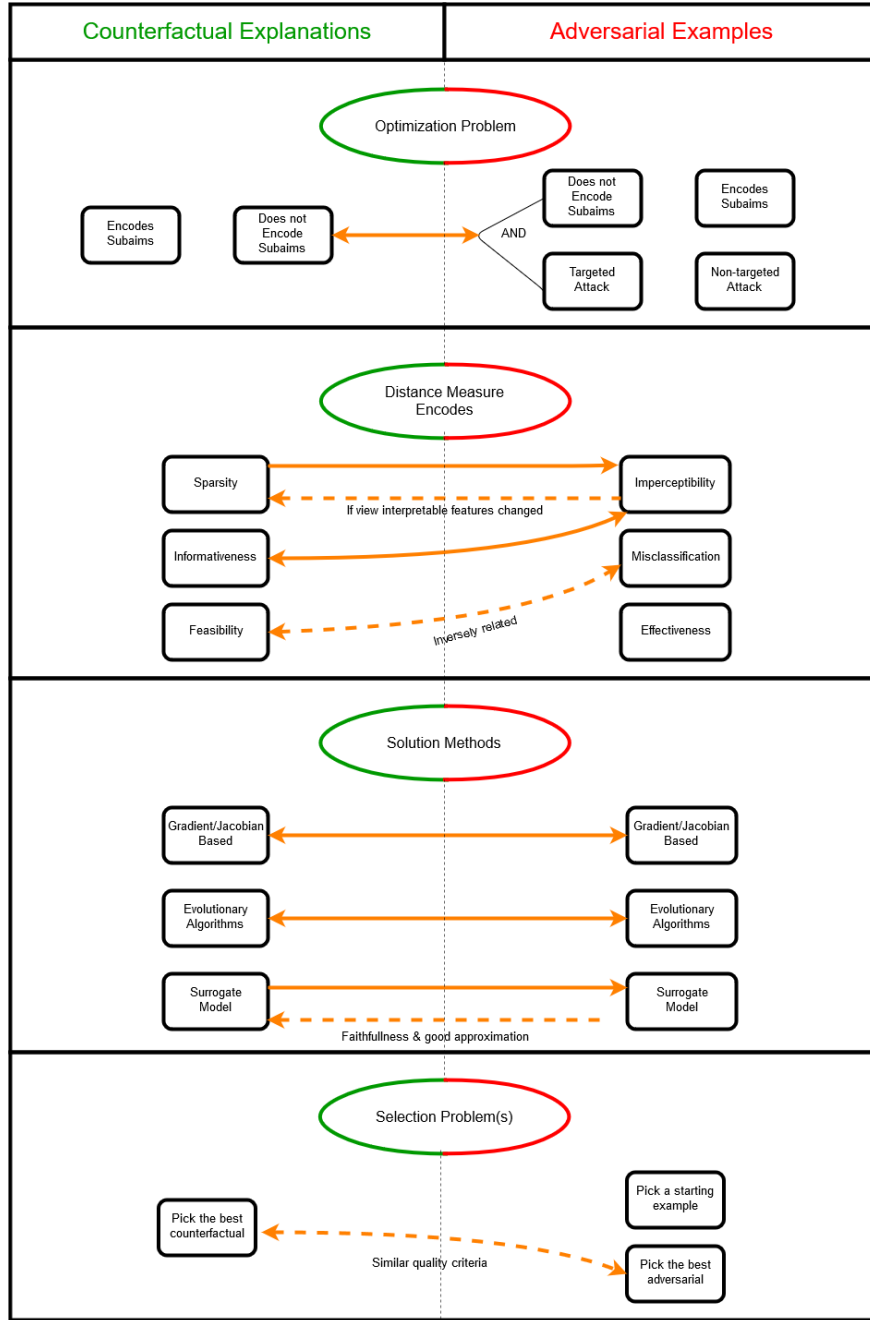
Figure 2: On the left-hand side, you see the counterfactual realm and on the right-hand side the corresponding adversarial concepts. Solid arrows between two items mean that a transfer is allowed in that direction. Dashed arrows mean that a transfer is possible under additional conditions, specified below the arrows.

18

of AEs in e.g. images are not per se irrelevant for CEs. Sparsity of CEs is only important for changes in interpretable features. For non-interpretable features, distributed changes can often be described as sparse changes in more abstract interpretable features. This may allow for further transfers from AE to CE distance measures.

The CE aim of informativeness and the AE aim of imperceptibility can also align. This alignment is discussed further in Section 4. Changing unexpected but effective features will very often lead to changes that are imperceptible if we are not explicitly pointed to them. The same holds vice versa. Hence, fruitful transfers can be expected. Feasibility in CEs requires that the alternative data point generated can realistically be reached by the explainee. Realistic data-points are those that are generally well represented in the training data. Hence the algorithm usually performs well in such cases and thereby feasibility counteracts the goal of misclassification. However, distance measures that have feasibility encoded could potentially be reversed to aid the aim of 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. This parallelism can be observed in the development of the fields. Both started with gradient-based methods, proceeded with evolutionary algorithms, and then considered surrogate models.[32] If applicable, solution methods developed for CEs can be easily used to generate AEs. The opposite direction might be more problematic. For CEs, approximately good solutions might not be good enough as they lead to bad/misleading explanations. This problem becomes particularly clear when we look at surrogate model approaches that are highly popular among AE researchers. 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, both face the selection problem among potential final AEs/CEs and it is an option to not only select one but several CEs/AEs. However, when generating AEs, the selection of the initial input is open, whereas in the case of CEs the selection is determined by the end-user. Moreover, the second selection problem also differs. First, unlike in CEs the solution space to non-targeted AEs contains vectors from different classes. Second, the number of CEs we can show is limited by humans' capacity to process information, while the number of AEs we can try in deceiving a system is generally unlimited. And even if we were to try to limit the number of vectors selected, it is highly unclear how solving the problem on the one hand would help solve the problem on the other. We would

---

[32]Interestingly, the areas concentrate on different solution methods. While in the literature on AEs mainly white box solvers are discussed, the literature on CEs mainly deals with black-box solvers. This is unexpected since AEs are usually considered from the perspective of an attacker without access to the model, while CEs are often built by the model engineers. A look at the use cases explains this paradox. Given 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.

need clear criteria as to what constitutes the better CE/AE, but the respective quality criteria and potential ranking functions might vary.

4 The Two Types of CEs

Until now we have discussed the similarities, differences and possible transfers between CEs and AEs. However, we have so far 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 will present a conceptual division into two types of CEs. We call them *feasible CEs* and *contesting CEs*. Feasible CEs are reasonable explanations from which recommendations for future actions can be derived. Contesting CEs are those that provide a basis for challenging an automated decision. AEs are very closely linked to the latter. We believe that this clarification may clear up some existing misunderstandings regarding the relationship between CEs and AEs. To clarify these issues, we will look at different scenarios that are manifested in the real world with different causal structures.

For all the presented scenarios we presuppose the following:

- There is a supervised machine 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. This machine learning algorithm is trained on a training set.

- There is an actual causal relation between variables. What does this mean? The input space I and the output space O of the algorithm f are defined as Cartesian products of sets of features $\mathcal{I}$ and $\mathcal{O}$. Each of these features relates to or is derived by some real-world properties. The real-world objects have specific causal relations, for which exists a true causal graph C. Often relevant features to complete the causal picture are missing. There are two ways to deal with such cases. One is to allow for latent variables Pearl [2009]. Option two is to ignore the incompleteness of the feature selection.[33] Since the latter is what is usually done in supervised learning contexts, we follow that procedure. 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.[34]

  We say $F \in \mathcal{I}$ is a causally irrelevant feature for $T \in \mathcal{O}$ if neither of the two conditions is met.

---

[33]Some features might even be inaccessible.

[34]Causal relevance in the first condition is clear. In the second condition, we allow the algorithm to assume that a change in the correlated feature F to T was due to a change in the variable L.

- In decision making, humans pay mostly attention to variables they consider relevant for the task [Jehee et al., 2011]. 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 actual 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 also a necessary condition for having many dogs (which are 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.
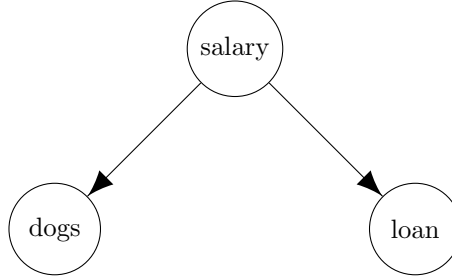
Figure 3: The causal graph contains three variables, salary, the number of dogs, and a binary variable for the loan application status.

Both types of CEs we introduce here indicate the features that the algorithm considered 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. Mixed type CEs where both causally relevant and irrelevant features are permuted are also possible and will be discussed.

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. Good CEs on the other side do exist.[35]

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, and given a certain threshold t is reached 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 would indeed be a good CE because it aids understanding and guides future actions. This is exactly what feasible CEs aim at. Feasible CEs 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 for instance over-/under-fitting, biased/lacking training data, missing features, etc [Bishop, 2006]. The second kind of reason is more principled. 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 have impact on 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 the 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

---

[35]This again makes clear that the class of CEs is broader than of AE as shown in the Appendix A + B.

the bank is a single-layer decision tree. Then, the algorithm may have 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 one dog, 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 reason for the algorithm's decision would be that threshold two for the number of dogs was not reached. A CE, in this case, would be:

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

This 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, would allow her to contest the decision, and in case she really urges for money she could use this information to deceive the algorithm. This is exactly how contesting CEs are characterized. Interestingly, an AE would be described by the same vector $(s, 2)$ and could potentially have the very same function, namely deceiving the system.

4.3   Mixed CEs

Usually, we do neither deal with the perfect algorithms from Section 4.1 nor the terrible algorithms of Section 4.2. Instead, we often have pretty good algorithms that mostly focus on relevant features to the right amount but sometimes make misclassifications. In such scenarios, we potentially have feasible, contesting, and also mixed CEs in which causally relevant and causally irrelevant features are perturbed.

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 1 fallacies have been avoided. As a result, the algorithm has learned, based on the training data, 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. The applicant has no dogs. In accordance with the ground truth, the loan application gets rejected since s is below t. However, the algorithm is not perfect. It learned that additional 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. Similarly 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 is not imperceptible. It changes the salary feature which as discussed above is not imperceptible since it is causally relevant.

4.4 AEs as Contesting CEs

This raises the question of whether every good AE makes a good contesting CE and vice versa. Indeed, the two classes have a great 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 point to the changes that were made. AEs on the other hand try to hide the changes as well as possible.
It is clear that every contesting CE is an AE as it must be misclassified to contest the decision by justified reasons. And, if there is a misclassification, there will potentially be contexts in which this bug can be exploited. Hence, 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 can happen 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. Especially 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 therefore argue that a minor change in all features means a change in only one interpretable feature, namely the coloration of the image. The change of coloration as the only feature is in fact a sparse change and it can rightly be argued that it is not causally relevant for classification. Thus, contesting CEs and AEs have at least a large 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 action-ability for generating counterfactuals. This was often achieved by the incorporation of 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 decision of the learning algorithm.

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 brings about that only these features are changed since we minimize the distance to the alternative input. Good 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 decisions of algorithms or deceive them. Thus contesting CEs work just like AEs.

Can this idea of the two complementary approaches be transferred from tabular data to images? We think that this is indeed 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 therefore potentially correlated to a different classification however it is causally irrelevant. This relates to the idea of Ilyas et al. [2019], where they discuss features that are predictive but not robust. 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. However, we have also shown where the two fields differ.

Can our analysis now shed some light on the three misconceptions we 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.[36] The most essential difference is that AEs, by definition, need to 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

---

[36] An example of a CE that is not an AE can be found in the Appendix B

that contesting CEs are another type of CEs that can be distinguished 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 great similarity with AEs. The third misconception concerns unjustified or missing transfers between the fields. We discussed under which conditions fruitful interactions are possible. While transfers on the level of the optimization problem or the solution methods are mostly permissible, transfers on the respective notions of distance are more demanding. In particular, 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, a preference based selection of feasible and contesting CEs can be developed. Also, various degrees of contestability could be introduced. Second, as suggested a large number of concepts from AEs can be transferred and used in generating CEs. Especially if the domains of the field show a greater overlap e.g. AEs for tabular data and CEs for image/audio data, transfers can be extremely beneficial. Conceptually we need further research on the relation between the paradigm of supervised learning and the transferability of AEs.

Acknowledgements

References

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

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

A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

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

E. R. Balda, A. Behboodi, and R. Mathar. Perturbation analysis of learning algorithms: generation of adversarial examples from classification to regression. *IEEE Transactions on Signal Processing*, 67(23):6078–6091, 2019.

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

S. Barocas, A. D. Selbst, and M. Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 80–89, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372830. URL https://doi.org/10.1145/3351095.3372830.

M. Beaney. Analysis. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2018 edition, 2018.

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

C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

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

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

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

R. Carnap. *Der logische aufbau der welt*, volume 514. Felix Meiner Verlag, 1998.

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

P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh. 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*, pages 15–26, 2017.

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

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

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

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

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

A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015. doi: 10.1080/10618600.2014.907095.

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

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

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

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

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

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

C. I. Ioannou, E. Pereda, J. P. Lindsen, and J. Bhattacharya. Electrical brain responses to an auditory illusion and the impact of musical expertise. *PLoS One*, 10(6), 2015.

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

J. F. Jehee, D. K. Brady, and F. Tong. Attention improves encoding of task-relevant features in the human visual cortex. *Journal of Neuroscience*, 31(22): 8210–8219, 2011.

D. Kahneman, S. P. Slovic, P. Slovic, and A. Tversky. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982.

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

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 2801–2807. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/388. URL https://doi.org/10.24963/ijcai.2019/388.

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

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

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

P. Menzies and H. Beebee. Counterfactual theories of causation. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2019 edition, 2019.

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

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

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

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

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

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

V. Navalpakkam and L. Itti. Modeling the influence of task on attention. *Vision research*, 45(2):205–231, 2005.

A. Páez. The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3):441–459, 2019.

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

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

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

J. Pearl. *Causality*. Cambridge university press, 2009.

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

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

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

A. Reutlinger. Extending the counterfactual theory of explanation. *Explanation beyond causation*, pages 74–95, 2018.

M. T. Ribeiro, S. Singh, and C. 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*, pages 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778.

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

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

B. Russell. On denoting. *Mind*, 14(56):479–493, 1905.

C. Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 20–28, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287569. URL https://doi.org/10.1145/3287560.3287569.

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

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

S. Sharma, J. Henderson, and J. Ghosh. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint arXiv:1905.07857*, 2019.

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

W. Starr. Counterfactuals. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2019 edition, 2019.

E. Štrumbelj and I. Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014. doi: 10.1007/s10115-013-0679-x.

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

J. Su, D. V. Vargas, and K. Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, Oct 2019. ISSN 1941-0026. doi: 10.1109/tevc.2019.2890858. URL http://dx.doi.org/10.1109/TEVC.2019.2890858.

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

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

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

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

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

G. Wheeler. Bounded rationality. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2020 edition, 2020.

X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824, 2019.

A   Appendix: Formal Proof for AE⊆CE

Here 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, $z \in TA(x, y', \delta, \| \cdot \|)$.

$\square$

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 is not true if the environment given by ϵ differs between the (non-)targeted ϵ CEs and AEs.

B   Appendix: Example CE but not AE

In Figure 4, we look at an example where we have a CE but not an AE.[37] 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. Also, the alternative generated to the eight which is classified as a nine does look like a nine. This cannot be an AE because it is not a misclassification.

---

[37]Image from Van Looveren and Klaise [2019]

Figure 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.