AI & SOCIETY
https://doi.org/10.1007/s00146-023-01837-4

MAIN PAPER

Understanding via exemplification in XAI: how explaining image classification benefits from EXEMPLARS

Sara Mann[1]

Received: 14 April 2023 / Accepted: 4 December 2023

Abstract

Artificial intelligent (AI) systems that perform image classification tasks are being used to great success in many application contexts. However, many of these systems are opaque, even to experts. This lack of understanding can be problematic for ethical, legal, or practical reasons. The research field Explainable AI (XAI) has therefore developed several approaches to explain image classifiers. The hope is to bring about understanding, e.g., regarding why certain images are classified as belonging to a particular target class. Most of these approaches use visual explanations. Drawing on Elgin's work (True enough. MIT Press, Cambridge, 2017), I argue that analyzing what those explanations exemplify can help to assess their suitability for producing understanding. More specifically, I suggest to distinguish between two forms of examples according to their suitability for producing understanding. I call these forms SAMPLES and EXEMPLARS, respectively. SAMPLES are prone to misinterpretation and thus carry the risk of leading to misunderstanding. EXEMPLARS, by contrast, are intentionally designed or chosen to meet contextual requirements and to mitigate the risk of misinterpretation. They are thus preferable for bringing about understanding. By reviewing several XAI approaches directed at image classifiers, I show that most of them explain with SAMPLES. If my analysis is correct, it will be beneficial if such explainability methods use explanations that qualify as EXEMPLARS.

Keywords Explainable artificial intelligence · XAI · Explainability · Understanding · Exemplification · Image classification · Example · Explanation

1 Introduction

Artificial intelligent (AI) systems have proven to be efficient tools in countless application contexts. Some of the most prominent applications involve systems that perform image classification tasks. Image classification is the process of matching a previously unlabeled image with the correct class label from a set of predefined classes. This task turned out to be challenging, but lately, advances in deep learning have given image classifiers a huge boost in accuracy (Goodfellow et al. 2016; Szeliski 2022). A well-known image classification task is handwritten digit recognition: The goal is to assign the correct target class (e.g., '7') to an image showing a handwritten number. Some applications of image classification involve considerable risk. For instance, AI systems can be used to classify pneumonia in chest X-rays (Yadav and Jadhav 2019), and autonomous vehicles need to categorize the objects they encounter to safely navigate their environment (Fujiyoshi et al. 2019).

The highest performance is currently achieved by image classifiers based on deep neural networks. Many of these systems are epistemically opaque, even to experts (Beisbart 2021; Burrell 2016; Mann et al. 2023). Such a lack of understanding can lead to ethical problems (Mittelstadt et al. 2016), violate laws and regulations (Goodman and Flaxman 2017), or hinder epistemic and practical ends (Boge 2021; Krishnan and Wu 2017). Accordingly, achieving understanding of such systems is taken to enable the fulfillment of a large number of desiderata that different stakeholders may have with respect to these systems (Langer et al. 2021).

Against this backdrop, the field of Explainable Artificial Intelligence (XAI) aims to render AI systems understandable (Beisbart and Räz 2022; Fleisher 2022; Langer et al. 2021; Páez 2019). XAI research has yielded an enormous number of explainability methods (see Arrieta et al. 2020; Molnar

✉ Sara Mann
sara.mann@tu-dortmund.de

[1] Department of Philosophy and Political Sciences, TU Dortmund, Dortmund, Germany

Published online: 27 January 2024

2022; Speith 2022). Many of them aim to explain aspects of image classification by means of visual explanations.[1] But what makes these explanations suitable for producing understanding? I argue that the theory of exemplification as proposed by Catherine Elgin (2017) allows to answer this question.[2] I show that specifying what the explanations of these methods exemplify reveals how much understanding they can provide. To this end, I distinguish between two forms of examples according to their suitability for producing understanding. I call these forms SAMPLES and EXEMPLARS, respectively.

To illustrate the difference between SAMPLES and EXEMPLARS, imagine the following scenario: You wish to understand why birdwatchers classify certain birds as belonging to a particular species, but you have no previous understanding of the bird's appearance. First, you turn to a photograph of the bird. Unbeknownst to you, the bird in the photo is an atypical specimen, and it is surrounded by several other birds you do not recognize either. Next, you consult a field guide with an illustration of the bird. To serve its specific purpose, the field guide's illustration depicts a typical specimen of the species. Moreover, it points up its characteristic features and disregards unspecific ones. Obviously, the field guide's representation is more helpful than the photograph in understanding what distinguishes the species. This is because the illustration provides epistemic access to contextually relevant features by exemplifying them. In this paper, I argue that existing explainability methods for image classifiers tend to produce explanations that are more like the cluttered photograph than like the straightforward illustration in the field guide. Their explanations are often difficult to interpret, because it is unclear which of their features they exemplify. I call such epistemically disadvantageous examples SAMPLES. Instead of SAMPLES, such XAI approaches should produce explanations which are tailored to context and mitigate the risk of misinterpretation. That is, they should explain with what I call EXEMPLARS.

A note on terminology: Although the terms 'example', 'sample', and 'exemplar' are often used interchangeably, I will use the notions EXEMPLAR and SAMPLE exclusively in the sense just described. 'Example' will be employed as an umbrella term to describe any instance that exemplifies some of its features, this comprises both EXEMPLARS and SAMPLES. I thereby depart from Elgin's terminology, according to which an *exemplar* is 'anything that exemplifies' (Elgin 2017, p. 184).

The paper is structured as follows. In Sect. 2, I characterize exemplification and its role for understanding. In Sect. 3, I review XAI approaches explaining image classification through the lens of exemplification. In Sect. 4, I spell out the distinction between SAMPLES and EXEMPLARS, and characterize EXEMPLARS in more detail. In Sect. 5, I discuss three potential objections to my proposal. In Sect. 6, I conclude and outline avenues for further research.

2 Understanding via exemplification

Different people may aim to understand different aspects of image classification. One may aim to understand how image classification works, why an individual image was classified in a certain way, where in the model the presence of a particular feature of the image was detected, and so on (see Zednik 2021, for different explanation-seeking questions in XAI contexts). For clarity, I will focus on one particular epistemic end in this paper: understanding why certain images are classified as belonging to a particular target class by a given classifier (e.g., understanding why certain images are classified as belonging to the 'robin' class). I refer to this specific epistemic end as *understanding*$_{class}$ in what follows.

For the purpose of this paper, I assume that achieving understanding$_{class}$ requires an explanation that specifies a dependency relationship between certain image features and the target class. By 'target class' I mean an image class (e.g., robin images), not a class of objects (e.g., actual robins). Furthermore, explainability strives to elucidate the workings of AI systems. That is, understanding$_{class}$ is about understanding the actual classifications of a given classifier. Those can diverge from true class membership. Despite these constraints, I believe that my proposal is not restricted to the specific case I focus on. Specifying what an explanation exemplifies illuminates the understanding it can produce. This holds for various epistemic ends and kinds of explanatory information, in image classification contexts and beyond.

In Elgin's view, examples can bring about understanding because they provide epistemic access to certain features of what they are an example of. This section examines how examples achieve this. First, I characterize examples and exemplification (Sect. 2.1). Second, I show that examples are well-suited to bring about an important aspect of understanding, namely, the grasping of contextually relevant relationships (Sect. 2.2).

[1] Most XAI methods do not provide explanations in the strict sense (Páez 2019). Nor do the 'explanations' that I focus on in this paper fit most accounts of explanation because they are visualizations and thus non-propositional. For the sake of simplicity, I will set aside this issue and adopt the common usage of the term in computer science.

[2] The usefulness of exemplification for analyzing XAI approaches is widely overlooked in the literature. An exception is Páez (2019, pp. 448–449), although he does not develop this thought further. To the best of my knowledge, Elgin herself is not concerned with (X)AI.

2.1 Examples and exemplification

According to Elgin, exemplification is one of the primary modes of reference (along with denotation, see below). An example 'functions as a symbol[3]' that makes reference to some of the properties, patterns, or relations it instantiates (Elgin 2017, p. 184). Thus, an example is always an example *of something* to which it refers. An image is, for instance, an example of the target class 'robin'. Likewise, an image can be an example of a feature of robins, such as the specific red of a robin's breast and face. When an example exemplifies some of its features, it makes those features manifest, highlights them, points them up (Elgin 2017, p. 185). The example, as Elgin puts it, *imputes* exemplified features to what it is an example of (cf. Elgin 2017, p. 266). If the example is interpreted correctly, this allows the interpreter to impute those features to the example's referent as well (cf. Elgin 2017, p. 253). By making some of its features salient, the example provides epistemic access to features of its referent. That is, exemplification does not only emphasize features that were already salient, but is itself a source of salience (Elgin 2017, p. 187).

In what follows, I characterize exemplification in more detail. As a comprehensive depiction of Elgin's theory of exemplification remains beyond the scope of this paper, I limit myself to Elgin's exposition of exemplification in Elgin (2017), and focus on five aspects which I deem relevant for exemplification in XAI contexts: instantiation, reference, typification, selectivity, and interpretation.

2.1.1 Instantiation

Exemplification requires instantiation. To exemplify characteristic features of members of the target class 'robin', the photograph must instantiate those features—say, pixels corresponding to a reddish breast and face. This is where exemplification differs from denotation: Denotation is a matter of stipulation (Elgin 2017, pp. 253–254). Saying 'Let this pencil represent a robin' is sufficient to let the pencil denote the bird, but it is not sufficient to exemplify it (cf. Elgin 2017, p. 185). The pencil may, however, instantiate certain features of robins—such as the particular shade of a robin's breast—and thereby come to exemplify them. Thus, an example that exemplifies features of its referent *instantiates* those features.

2.1.2 Reference

Mere instantiation is not sufficient for exemplification. There are countless images of robins. If they do not refer to the target class 'robin', they are not examples of it (cf. Elgin 2017, p. 185). Consider an image showing a robin sitting on a birch tree. In one context, this image may refer to the target class 'robin'. However, in another context the same image may refer to the 'birch' class only, even though it still shows a robin. Clearly, in the latter case the image is not an example of the 'robin' class. Thus, an example that exemplifies features of its referent *refers* to those features. Reference is not difficult to achieve. For instance, labeling an image as belonging to a specific target class suffices to fix the reference (cf. Elgin 2017, p. 255). When an explanation is consulted in XAI, the reference is usually determined by the context. An explainability approach is used to answer a specific question, and the resulting explanation is analyzed accordingly. Therefore, I take reference for granted in the following.

2.1.3 Typification

What makes an example epistemically effective is that it points beyond itself. If a robin image exemplifies pixels corresponding to a reddish breast and face, it serves as an example that points to the extension of all and only instances sharing these features. In other words, the example typifies this extension. By typification, the example makes salient in what respect the members of this extension are similar, it likens them to one another (Elgin 2017, pp. 263–267). Thus, an example that exemplifies features of its referent *typifies* the extension of all and only instances that share these features.

2.1.4 Selectivity

Exemplification is also selective. As indicated above, exemplifying certain features involves highlighting those features. However, while an example can emphasize multiple of the features it instantiates, it cannot emphasize all of them at once (Elgin 2017, p. 185). Instead, highlighting or emphasizing certain features leads to omitting, downplaying, or sidelining others. The field guide's illustration of a robin may instantiate and emphasize a colored breast while instantiating but sidelining a particular pose or size. Selectivity is not necessarily problematic because an example need only exemplify relevant features to be effective. Which features count as relevant depends on the contextual function of the example (Elgin 2017, pp. 193–194). Thus, an example that exemplifies features of its referent *selectively* highlights those features, while sidelining or omitting other features that the example instantiates.

[3] Elgin has a broad concept of symbol that comprises both linguistic and non-linguistic objects, including works of art or scientific models.

2.1.5 Interpretation

To produce understanding, an example needs to be interpreted correctly in at least two ways. Both relate to specifying the extension an example typifies. The first way of interpreting an example involves recognizing which of the features it instantiates are exemplified features (Elgin 2017, p. 188). This can be tricky, because exemplification is selective. Not all the features an instance instantiates are also exemplified. Elgin distinguishes between inert, scaffolding, and exemplified features (Elgin 2017, pp. 265–266). *Inert features* are simply irrelevant, e.g., the typeface the class label is in. *Scaffolding features* are not themselves exemplified, but enable exemplifying. For instance, a field guide illustration may contain small arrows to point out characteristic features of a species (as in Peterson 1980). Finally, *exemplified features* are those which are imputed to the referent. Correctly interpreting an example requires identifying which of its features are exemplified, and thus which extension of instances the example typifies. To illustrate, an example of the target class 'robin' usually instantiates innumerable features that are not characteristic of members of that class. Suppose the correct interpretation of this example is to identify as exemplified feature a reddish breast. This allows identifying the extension that the example typifies, namely all instances that share this feature.

What Elgin calls *stage setting* can help identify exemplified features by reducing the ambiguity of the example (Elgin 2017, p. 192). Stage setting can be performed with any example by providing additional information. For instance, an ornithologist may point out characteristic features of a bird. Even if the bird is not a typical specimen, this may be enough to ensure correct interpretation. Also relevant contrasts can facilitate interpretation (Elgin 2017, pp. 188–192). Comparing a robin to a red-breasted flycatcher may allow one to infer its distinctive features. Stage setting can also be accomplished by using an example that is easier to interpret from the start. One may choose a typical instance that possesses the relevant features to a high degree, and where these features are prominent and easy to discern (Elgin 2017, p. 192). It can be even useful to *create* an example to make certain features more salient from the outset. Elgin's standard example is a paint company's sample card that can be used to select a color for a paint job (Elgin 2017, pp. 187–188). The sample card instantiates only those features of the paint that are to be exemplified (i.e., the color shade). Together with background knowledge about sample cards, this ensures correct interpretation.

Examples need to be interpreted in a second way. Interpreting an example also involves recognizing what the extension it typifies amounts to. An example typifies the extension of instances that share the features it exemplifies. But which extension is that (Elgin 2017, p. 190)? Say, a robin image typifies the extension of all and only images that possess pixels corresponding to a reddish breast and face. To make use of this information, one needs to know whether this is the extension of typical robin images, of all robin images, or only of a minority. This information is usually not provided by the example itself. Which extension it represents depends on the way it was chosen or generated (Elgin 2017, p. 190). In the context of explaining image classification, this information needs to be provided by the XAI method in question.

In these two ways, an example that exemplifies features of its referent needs to be *interpreted* correctly to be used to project to the extension of instances that share those features.

Instantiation, reference, typification, selectivity, and the need for interpretation are characteristics all examples have in common. However, examples do not always yield the desired epistemic outcome. To begin with, examples can be misleading. They can purport to exemplify features of their referent that the referent does not actually possess (Elgin 2017, p. 199). An image of a robin that imputes the feature of a blue breast and face to robins purports to exemplify a feature that robins actually do not share. Examples can also be interpreted incorrectly (Elgin 2017, p. 189). For instance, the interpreter can project features of the example to the example's referent that the referent does not have. The user of a field guide who interprets a robin illustration as exemplifying the two-dimensionality of robins misinterprets the example. In other cases, the interpreter misconstrues the example's referent (cf. Elgin 2017, p. 190). A birder might think that a particular bird is a robin, but actually, it is a redstart. Thus, not all examples are good examples, and not all interpreters are competent.

After having characterized exemplification in general, I return to the particular case I focus on in this paper. Which examples are suitable to bring about understanding$_{class}$? To begin with, such an example takes the form of a visual explanation provided by an XAI approach. It needs to exemplify features that a given classifier uses to classify images as belonging to the target class in question. For instance, if a classifier classifies images that possess pixels corresponding to a reddish breast and face as members of the 'robin' class, the example should exemplify this feature. Unfortunately, things are usually more complicated. Most modern image classifiers rely on deep neural network architectures (Szeliski 2022), such as *Deep Convolutional Neural Networks* (DCNN; see LeCun and Bengio 1995) or *Vision Transformers* (Dosovitskiy et al. 2021). Such systems are not coded by hand, but rely on complex sub-symbolic representations (Smolensky 1988) that develop during the training phase. These representations enable the system to exploit a large number of different image features. It likely exceeds the cognitive capabilities of humans to grasp such a complex representation in its entirety. Certainly, it cannot be cast in terms of a neat example.

Instead of completely abandoning the project of understanding modern image classifiers, one could set more modest goals. An example likely cannot exemplify all the features a given classifier relies on for classification. But it can provide partial epistemic access by exemplifying some of them. Furthermore, the amount and diversity of exemplified features can be increased by using multiple examples. A single example of the 'robin' class cannot exemplify pixels corresponding to upperparts that are *either* gray *or* olive *or* brown, but three different examples can. However, increasing the number of examples to include more features becomes infeasible at some point. In many cases, even dozens of examples are not enough to express all the features that a classifier has learned. Therefore, we need to draw on context to make a selection. To achieve understanding$_{class}$, obvious candidates are the most influential features. Even though a classifier may be able to classify robin images as 'robin' when the reddish face is occluded by leaves, the reddish face may still be among the most typical features. Other contexts, however, may require to gain epistemic access to less influential features, or to a particular feature the classifier exploits. Thus, it is contextually relevant features that need to be exemplified. This can be achieved by one or more examples of the target class. For simplicity, I will assume that producing understanding$_{class}$ requires one or more examples that exemplify features the classifier typically relies on to assign images to a given target class.

To sum up: To produce understanding$_{class}$, one or several visual explanations serve as examples of the target class in question. Such an example exemplifies the most influential features the given classifier relies on for classification. It *instantiates* these features, *refers* to them, *typifies* the extension of all and only instances that share those features, *selectively* highlights them, and needs to be *interpreted* correctly. Before returning to understanding$_{class}$ in particular, I show how examples can bring about understanding more generally.

2.2 Understanding

According to Elgin, an example that exemplifies features of its referent can afford an understanding of its referent because it makes contextually relevant features of the latter salient (Elgin 2017, p. 249). She complements this idea by a thorough account of understanding. However, I want to offer another way of linking examples to understanding. I suggest that the epistemic access examples provide can be spelled out in terms of understanding-related abilities. More specifically, I think that understanding$_{class}$ involves the ability to judge whether a given instance belongs to that class. This is supported by existing literature on the subject.

It is a common view that understanding and abilities are closely linked (Baumberger et al. 2017). Also (Elgin 2017, p. 3) thinks that understanding 'involves being able to draw inferences, raise questions, frame potentially fruitful inquiries, and so forth.' The motivation for linking understanding to abilities is the thought that understanding involves 'grasping' dependency relations, or 'seeing how things hang together' (e.g., Elgin 2017; Hills 2016; Kvanvig 2003; Riggs 2003). This grasping, then, is often conceptualized in terms of specific (mostly cognitive) abilities.

Although further understanding-related abilities are discussed in the literature (see, e.g., de Regt 2015; Khalifa 2017; Newman 2017; Strevens 2013), I want to focus on a specific ability: the ability to reason not only about an individual instance, but also about similar or hypothetical cases. Such an ability is stressed by several authors. For instance, Hills (2016) characterizes understanding-why in terms of specific abilities that involve drawing explanatory inferences about current and similar cases. Similarly, Grimm (2011, p. 89) claims that grasping comes with 'a modal sense or ability' to draw inferences about counterfactual cases. This thought can also be found in Woodward's suggestion that understanding is related to the ability to answer what-if-things-had-been-different questions (Woodward 2004).

I suggest that the ability to reason about similar or hypothetical cases can be gained through effective examples. An example typifies the extension of all and only instances that share the features that it exemplifies. Recognizing this extension allows to judge whether a given instance belongs to this extension. In other words, it allows to draw inferences about instances that diverge from the example. Although understanding requires more than grasping (Baumberger 2019), I take it that examples can bring about an important aspect of understanding.

In Sect. 2.1, it became clear that it is unlikely that an example can provide epistemic access to all features that a given classifier relies on for a particular classification. Accordingly, it is unlikely that an example can typify the extension of all and only instances that are classified as belonging to the target class in question. This means that an example may not afford the ability to draw inferences about all instances of a target class, and may not provide a perfect understanding$_{class}$. But understanding, and also grasping, comes in degrees (Baumberger et al. 2017; Baumberger 2019). Even if it is impossible to gain a complete understanding of the inner workings of modern image classifiers, one can gain a partial understanding. Even if it is impossible to decide for every instance whether it belongs to the extension of target class members, one can gain the ability to judge a relevant portion of instances. And if that is insufficient, one can use multiple examples to extend one's understanding.

Again, the required degree of understanding depends on context. For instance, a contextually sufficient degree of understanding$_{class}$ may involve the ability to judge typical cases, but not outliers. In other cases, all that is needed may be the ability to predict whether a particular feature usually yields the classification in question. Context determines what needs to be understood, and this determines what an example needs to exemplify. If an example meets the requirements context provides, it is a powerful tool for achieving understanding.

3 Do XAI approaches provide effective examples?

There is a vast amount of XAI approaches that aim to explain image classification by providing visual explanations (see, e.g., Molnar 2022). In this section, I show that specifying what such an explanation exemplifies elucidates its potential for bringing about understanding—and also the risk of misunderstanding it bears. The guiding question is whether an approach provides examples which can bring about understanding$_{class}$, that is, which enable to understand why a given system classifies certain images as belonging to a particular class. What exactly this entails may vary with context. As indicated in Sect. 2.1, I will assume that producing understanding$_{class}$ requires one or more examples that exemplify features the classifier typically relies on to assign images to a given target class. My analysis reveals that we are currently lacking XAI methods that harness the full potential of examples for producing understanding.

Below, I review a selection of explainability methods that are directed at image classifiers. The review is not restricted to methods that are categorized as 'explanation by example' in taxonomies of XAI methods (see Arrieta et al. 2020; Belle and Papantonis 2021; Lipton 2018; McDermid et al. 2021; Schwalbe and Finzel 2023; Speith 2022). Also explanations generated by other common methods can be seen as producing understanding by exemplifying some of their features. I grouped the methods into three categories: First, methods that adduce several actual members of a target class (Sect. 3.1); second, methods that highlight influential areas within an actual class member (Sect. 3.2); and third, methods that create an artificial image by visualizing features characteristic of a target class (Sect. 3.3).

3.1 Approaches that adduce several actual class members

A straightforward way to exemplify is to pick an instance which instantiates the features that are to be exemplified. There are several XAI methods that take such an approach. In most cases, they provide not only one, but multiple examples. As discussed in Sect. 2.1, this allows to exemplify more features that are characteristic of target class members, and thus to paint a more complex picture of the classifier's representation of that class. However, as we will see below, such methods lack sufficient stage setting to ensure correct interpretation.

The method proposed by Bien and Tibshirani (2011) selects a number of so-called *prototypes*. A prototype set consists of several data instances that are taken to be representative of clusters in the data distribution of a target class, and that are therefore suited to 'graphically summarize' (Bien and Tibshirani 2011, p. 2418) its elements. The *MMD-critic* method by Kim et al. (2016) provides not only prototypes, but complements them with so-called *criticisms*. Criticisms are instances that are part of the data distribution, but not well represented by the prototypes. Together with prototypes, criticisms aim to reflect the complexity of the data set. Another way to identify typical instances of a target class is to choose input images that had a large impact on the model parameters or predictions during training (so-called *influential instances*; see Koh and Liang 2017). Finally, *dataset examples* are actual images taken from the data set that maximally activate a specific unit of a neural network (Schubert et al. 2021). In terms of exemplification, the above methods provide multiple examples that aim to exemplify features that the classifier typically relies on for classification.

Indeed, it is plausible that these methods provide examples that *instantiate* those features. But instantiation is not sufficient. Achieving understanding$_{class}$ requires to correctly interpret the examples. Recall that interpretation comes in at least two ways. On the one hand, it involves recognizing what the typified extension amounts to. The above methods make this relatively clear: They provide examples that aim to typify the extension of typical members of a target class. On the other hand, interpretation involves to recognize exemplified features. However, explainability approaches that adduce several examples of a class do not provide much stage setting to facilitate this. Their examples are actual data instances that may instantiate a plethora of features. The person interpreting the examples must decide which of its features each example exemplifies. This can go wrong, since AI systems regularly detect different patterns than humans do (Mueller 2020). A human confronted with a number of robin images draws on her world knowledge and probably assumes that certain features of the bird are exemplified by the examples. However, it may be that the model learned a representation of the target class that does not match the human conception of robins, but responds to watermarks, backgrounds, metadata, or other correlated properties the human does not even notice (see Lapuschkin et al. 2019; Ribeiro et al. 2016). A failure to identify the exemplified features entails a failure to project to the extension of instances the examples typify. In such cases, the examples do not afford the ability to judge whether a new instance belongs to this extension: They fail to produce understanding$_{class}$.

In sum, methods that adduce several examples of a target class provide explanations that instantiate inert features that may be mistaken for exemplified ones. Also, these approaches do not provide much stage setting to ensure correct interpretation. This makes them instances of what I call SAMPLES. Still, these methods may work sufficiently well in many cases. My point is that their examples are not ideal. In the following section, I discuss explanations that aim to make explicit which of the features they instantiate are the exemplified ones.

3.2 Approaches that highlight influential areas within an actual class member

Pixel attribution-based methods are probably the best-known family of explainability approaches for image classification. Such methods emphasize those areas of an input image that were most relevant to its classification. There exist numerous methods that share this idea (for an overview, see Molnar 2022; Schwalbe and Finzel 2023). Many approaches generate *saliency maps* which visualize the influence that individual pixels or groups of pixels (‘super-pixels’) had on the output class. Some methods produce explanations by altering the input image. For instance, influential regions can be highlighted by adding a colored overlay (e.g, Petsiuk et al. 2021) or by masking less influential areas (e.g., Ribeiro et al. 2016). In terms of exemplification, such approaches introduce scaffolding features (i.e., colored or masked areas) into the image. Other methods set the stage by using the saliency map as an addendum to the original image (e.g., Fong and Vedaldi 2017; Lapuschkin et al. 2019). In both cases, saliency maps make explicit which image areas instantiate exemplified features, and which instantiate inert ones.

Being so-called *local* methods, pixel attribution approaches are not intended to produce understanding$_{class}$. They do not aim to explain the classification of a wider range of images, but of a single instance. As a result, the features saliency maps exemplify cannot be used to project to the extension of all instances that share these features (cf. Alvarez-Melis and Jaakkola 2018; Lipton 2018). Their scope is much more narrow, and often difficult to determine (see, e.g., Ribeiro et al. 2018, p. 1528). This hampers both ways of interpretation: It remains unclear what extension the examples provided by local methods typify, and what this extension amounts to.

The *Anchors* method (Ribeiro et al. 2018) aims to address this limitation. Anchors generates local explanations consisting of if-then rules. These rules ‘anchor’ the prediction because they hold for a specific range of instances. An anchor consists in one or more feature values. If the anchor is present in a given data instance, it will be classified in a certain way with a high probability. Thus, Anchors differs from the above-mentioned methods because its examples typify the extension of instances that share the anchored features, and affords the ability to judge new cases. Furthermore, the method allows to generate multiple anchors that cover the classifier’s global behavior (Ribeiro et al. 2018, p. 1533). This facilitates to assess what the typified extension amounts to, and makes Anchors a candidate for gaining understanding$_{class}$. However, the interpreter must synthesize the information provided by multiple anchors to achieve a more global understanding of the classifier’s representation. As seen in the previous section, this increases the risk of misinterpretation.

Methods based on pixel attribution have further limits. First, there are features which they cannot visualize because they operate at the pixel level only.[4] This concerns general image attributes such as contrast or brightness (Alqaraawi et al. 2020, p. 283), but also higher-level features the classifier may have learned (e.g., a specific part of an object). Clearly, if saliency maps cannot visualize these features, they cannot help to exemplify them either. Second, several experiments have shown that some approaches are not sufficiently sensitive to the classifier’s behavior and may produce misleading explanations (e.g., Adebayo et al. 2018; Ghorbani et al. 2019; Gu and Tresp 2019; Wilking et al. 2022). In such cases, explanations purport to exemplify features associated with a target class, but actually they fail to do so. Third, in spite of the stage setting saliency maps provide, it may not always be clear what the exemplified features are. Saliency maps make image *regions* salient, not image features. These regions may instantiate inert features as well as exemplified ones. To illustrate with Anchors, the anchors used to explain image classification are super-pixels taken from the explained instance. Oftentimes, it is not obvious how these image patches can be mapped to new instances, as it is not clear how close the new instance must match the anchor (Molnar 2022, p. 214). This moves these examples in the direction of SAMPLES.

In sum, examples generated by pixel attribution have the advantage of involving more stage setting than prototype-based methods. On the other hand, the amount of stage setting may not always suffice to ensure correct interpretation, the examples cannot visualize all potential classification criteria, and some explanations may not be robust. Nevertheless, these methods are suitable for providing understanding in certain contexts. Local methods can afford the ability to explain the classification of current and maybe certain similar instances, provided that the explainability approach states the explanation’s scope. However, local methods are not designed to produce understanding$_{class}$. This is different with the methods I consider in the next section.

[4] I thank an anonymous reviewer for highlighting this problem.

3.3 Approaches that create an artificial image

As discussed in Sect. 2.1, achieving understanding$_{class}$ requires an example that exemplifies features which the classifier typically relies on to classify images in a particular way. Those features are stored in the classifier's representation of a target class. The previously discussed methods are capable of providing epistemic access to some of these features. However, to produce understanding$_{class}$, it makes sense to directly visualize the classifier's representation. This combines the advantages of prototype-based methods and saliency maps. On the one hand, directly visualizing the representation removes contingent features that prototypes possess. On the other, this produces an example that allows to project to a larger portion of target class members than local methods do.

The representations of neural networks can be visualized by *feature visualization* (see Olah et al., 2017, for an overview). The underlying principle is called *activation maximization* (Erhan et al. 2009). Roughly put, this method generates one or more images to which some unit of a classifier maximally reacts. Thus, this approach is fundamentally different from the methods discussed previously, because the latter use actual data for their explanations. Feature visualization can be performed for any unit of the network, ranging from individual neurons to the final probability of a class (Molnar 2022, p. 244). To achieve understanding$_{class}$, it is useful to generate images that maximize the final classification probability. In terms of exemplification, this means to generate examples that mainly instantiate features that are associated with the classifier's representation of a class. This facilitates both ways of interpretation: Like the methods I discussed in Sect. 3.1, the approach makes explicit what the typified extension amounts to (i.e., typical members of the target class in question). Furthermore, activation maximization reduces the instantiation of inert features. This helps to identify those features that are imputed to the referent. As mentioned in Sect. 2.1, a single example cannot exemplify all features that are associated with a target class. However, generating multiple feature visualizations allows to exemplify more of them.

There are different ways and degrees of constraining the generated images. By imposing relatively few constraints, one obtains surrealistic visualizations that depict multiple, partial, and perspectively distorted instances belonging to the represented class (see Mordvintsev et al. 2015; Simonyan et al. 2014; Yosinski et al. 2015). This hampers correct interpretation, since it is difficult to project from such examples to actual instances. Applying stronger constraints leads to more or less natural-looking images that normally depict a single, complete instance (see Nguyen et al. 2016, 2017). Such images can be used more easily to project to natural instances. The downside of stronger constraints is that they introduce correlated features into the image (Olah et al. 2017, n. p.). Again, it is uncertain which features are exemplified and which are inert.

Despite these difficulties, I take it that feature visualization is better suited to produce understanding$_{class}$ than the methods that I discussed in the previous sections. The examples generated by feature visualization are easier to interpret because they instantiate less inert features, and they are designed to provide epistemic access to features associated with an entire target class. When the aim is to achieve understanding$_{class}$, these examples are already close to what I call EXEMPLARS. In the following section, I glean the previous insights and characterize EXEMPLARS in more detail.

4 How explaining image classification benefits from EXEMPLARS

As I hope to have shown, examples which exemplify contextually relevant features are a powerful tool for achieving understanding in image classification contexts. In this section, I introduce EXEMPLARS as a type of example that is especially useful for producing understanding. In Sect. 4.1, I spell out the distinction between EXEMPLARS and SAMPLES. In Sect. 4.2, I point out five desirable properties of EXEMPLARS that make them epistemically effective.

4.1 EXEMPLARS VS. SAMPLES

My review of different XAI methods in Sect. 3 showed that all of them are capable of providing epistemic access to features that they share with their referent. However, it also revealed that some examples are easier to interpret than others. Maybe surprisingly, it is not necessarily 'realistic' images that are best suited to produce understanding. In most cases, natural instances of a target class instantiate numerous contingent features that are not relevant to the question at hand. These inert features can be mistaken for exemplified ones and can lead to misunderstanding. By contrast, an artificial image allows for extensive stage setting to facilitate interpretation.

In Sect. 1, I sketched the idea of EXEMPLARS by referring to the illustrations in field guides. In line with my proposal, Law and Lynch (1988) argue that a schematic illustration of a given bird species is better suited to highlight its relevant features than a photograph, as the former 'provides clear criteria by artfully rendering most possible differences irrelevant' (Law and Lynch 1988, pp. 284–285). The schematic illustrations that the authors refer to are by Roger Tory Peterson, an ornithologist whose field guides have been formative in the field. Peterson was a proponent of using schematic illustrations for bird identification. He thought that they were better suited to emphasize *field marks*, i.e. characteristic features of a species:

> A photograph is a record of a fleeting instant; a drawing is a composite of the artist's experience. The artist can edit out, show field marks to best advantage, and delete unnecessary clutter. [...] A photograph is subject to the vagaries of color temperature, make of film, time of day, angle of view, skill of the photographer, and just plain luck. [...] Whereas a photograph can have a living immediacy, a good drawing is really more instructive. (Peterson 1980, pp. 9–10)

Like photographs of birds, SAMPLES may contain 'unnecessary clutter' or lack contextually relevant features. This makes it difficult to decide which features are of interest. At best, SAMPLES complicate the acquisition of understanding because their interpretation requires considerable effort. At worst, SAMPLES inhibit the acquisition of understanding because they are misinterpreted or fail to exemplify contextually relevant features. By contrast, I call examples that mitigate the risk of misinterpretation by sufficient stage setting EXEMPLARS. EXEMPLARS are intentionally designed or chosen to produce the understanding that is required in a given context. They make explicit which of their features are relevant, be it through omission, highlighting, or other forms of stage setting. EXEMPLARS are clear cases that reduce cognitive effort and the risk of misunderstanding (cf. Elgin 2017, pp. 168, 192). In other words, they are like illustrations in field guides.

Here are some clarifications. First, the main difference between EXEMPLARS and SAMPLES is the amount of stage setting required to ensure correct interpretation. Often they do not form clearly delineated categories, but differ only in degree. Second, I do not claim that EXEMPLARS are necessarily artificial instances. In principle, also a natural image can be easy enough to interpret to be considered an EXEMPLAR. However, artificial examples allow for more stage setting. They may therefore be preferable in certain contexts. Finally, EXEMPLARS are not the right tool for all conceivable epistemic ends in image classification contexts. There may be aspects of image classification that cannot be visualized in terms of even several EXEMPLARS, such as classification criteria of large and heterogeneous target classes like 'animal' or 'fruit'. In other cases, EXEMPLARS may not be suitable for all addressees. For instance, a Google app classified photographs of Black people as 'gorilla'.[5] Although this information is crucial for improving the classifier, generating an EXEMPLAR visualizing such classification criteria can cause harm and may be unacceptable in certain explainability contexts.[6] Other EXEMPLARS are ethically and scientifically inadmissible independent of the addressee. For instance, Wu and Zhang (2016) merged photographs of convicted offenders. Making dubious physiognomic assumptions, they hoped to thereby gain epistemic access to 'subtypes of criminal faces' (Wu and Zhang 2016, p. 8).[7] If not precluded by such considerations, EXEMPLARS can be a valuable tool for achieving understanding in image classification contexts.

4.2 Desirable properties of EXEMPLARS

I suggest that EXEMPLARS can be characterized by five desirable properties: They should (1) point up exemplified features, (2) disregard inert features, (3) have a clear scope of application, and exemplify features that are both (4) contextually relevant and (5) intelligible. I discuss each aspect in turn.

4.2.1 Pointing up exemplified features

EXEMPLARS should point up exemplified features to distinguish them from inert ones. This facilitates interpretation and reduces the risk of misunderstanding. Pointing up exemplified features is a type of stage setting and can be achieved in different ways—e.g., by providing additional instructions, presenting relevant contrasts, or removing distracting features (Elgin 2017, p. 192).

Existing XAI approaches in image classification contexts choose different strategies to point up the features their explanations aim to exemplify. Prototype sets (Sect. 3.1) select typical class members. Saliency maps (Sect. 3.2) use scaffolding features or additional information to direct attention to relevant image features. Images generated by feature visualization (Sect. 3.3) try to instantiate mostly exemplified features. Further strategies include providing visual counterfactuals (e.g., Dhurandhar et al. 2018; White et al. 2021) or additional verbal explanations (Rabold et al. 2020).

Pointing up exemplified features can help detect unexpected behavior in AI systems. It is a well-known problem that image classifiers may rely on proxies (Lapuschkin et al. 2019) or 'shortcuts' (Geirhos et al. 2020) that do not generalize to new data. In many cases, these features are counterintuitive from a human viewpoint and thus easily overlooked. An EXEMPLAR that points up those proxy features can make the problem manifest.

[5] See, e.g., https://www.bbc.com/news/technology-33347866, accessed: 22 August 2023.

[6] I would like to thank Dick Timmer for raising this point.

[7] Thanks to Kevin Baum for drawing my attention to this (now withdrawn) paper.

4.2.2 Disregarding inert features

Exemplars should disregard inert features to distinguish them from exemplified ones and to thereby facilitate interpretation. Disregarding inert features and pointing up exemplified ones often go hand in hand. Thus, many of the above-mentioned strategies for highlighting exemplified features can be also seen as strategies for disregarding inert ones.

This is illustrated by existing XAI methods. Prototype sets (Sect. 3.1) exclude atypical instances, saliency maps (Sect. 3.2) help downplaying certain features by highlighting others, and feature visualization (Sect. 3.3) provides examples which omit most atypical features. Oftentimes, omission is the preferable strategy to disregard inert features. However, it may not always be possible to omit them entirely. For instance, instantiating color cannot be avoided (cf. Elgin 2017, p. 192). An image that shows a 'colorless' shape is indistinguishable from an image that shows, e.g., a white shape. Those cases require more stage setting, such as additional instructions for interpretation. Further features which can hardly be omitted are those that are subject to contingent variation, e.g., regarding size or viewing angle. These are cases where certain features can only be exemplified by instantiating others (e.g., the presence of a beak can be exemplified only by also instantiating a specific beak position). I return to this aspect in Sect. 5.

4.2.3 Clear scope of application

The scope of application of exemplars should be clear. While the two previous properties of exemplars facilitate interpretation by aiding the identification of exemplified features, knowing the scope of application helps to recognize what the typified extension amounts to. This is especially relevant in human-AI interaction, where users may anthropomorphize image classifiers and falsely attribute their own classification strategies to them (see Mueller 2020).

In some cases, the scope of an exemplar could be clarified by providing information about the classifier and its limitations. Consider a classifier used to detect skin cancer that was trained mostly on data from light-skinned people (Goyal et al. 2020; Wen et al. 2021). In such a scenario, users may need to know that an exemplar of the 'melanoma' class does not necessarily represent the behavior of the classifier with respect to data from people with darker skin. Also the XAI approach in question can impact the scope of an exemplar. For instance, it is important to be aware that the scope of local explanations like saliency maps (Sect. 3.2) is much more narrow than that of global explanations like those generated via feature visualization (Sect. 3.3). Indeed, as seen with Anchors (Ribeiro et al. 2018), some XAI approaches try to make the scope of their explanations explicit. Finally, the nature of exemplars influences their scope as well. Exemplars provide epistemic access to potentially complex referents by foregrounding some of their features. This may involve partial misrepresentations of their referents. Failure to recognize these limitations may lead to misunderstanding. As discussed in Sect. 2.1, an example can only provide epistemic access to some of the features that are associated with a given target class. In such cases, the inferences that an exemplar warrants may not be universally valid. Therefore, it may be necessary to ensure that the explanation is only used to project to instances that it represents sufficiently well.

4.2.4 Exemplification of contextually relevant features

Exemplars should exemplify only contextually relevant features. It is widely held in XAI research that context affects what kinds of explanations an explainability approach needs to deliver (Langer et al. 2021; Miller 2019; Nyrup and Robinson 2022; Páez 2019; Zhou et al. 2022). Candidates for relevant context factors include the explanation's addressee, the stakes, or time constraints (Langer et al. 2021). Furthermore, different contexts may involve different epistemic ends (see Zednik 2021). An exemplar should respond to such factors by exemplifying features to which epistemic access is needed in a given context.

A variety of features can be relevant in a given context. In this paper, I focused on the image features that were most influential to a given classification. However, XAI methods are able explain other aspects as well. For instance, the relative influence of different image regions can be exemplified by saliency maps that use different colors to express degrees of influence (see Petsiuk et al. 2021). Counterfactual dependencies may be exemplified by visual counterfactuals (e.g., Dhurandhar et al. 2018; White et al. 2021). Multiple exemplars could be used to visualize different 'classification strategies' of a classifier, like those identified by Lapuschkin et al. (2019). Whether image features are (roughly) sufficient for a classification can be expressed by the Anchors method (Ribeiro et al. 2018) discussed in Sect. 3.2. Also the influence of a particular feature may be contextually relevant. The TCAV method by Kim et al. (2018) allows to assess the influence of user-generated concepts (e.g., 'striped') on the final classification (e.g., 'zebra'). Finally, additional verbal explanations can be used to describe unvisualizable classification criteria (Rabold et al. 2020).

4.2.5 Exemplification of intelligible features

The features exemplified by exemplars should be humanly intelligible in the sense that they can be mapped to features of new instances. Grasping how image features and target class are related requires that these features make sense from a human point of view. These features can then be exemplified by an EXEMPLAR.

Features that are intelligible in this sense need not correspond to human concepts, and they need not be recognized as distinct features before they are exemplified (cf. Elgin 2017, p. 187). For instance, Schubert et al. (2021, n. p.) discovered 'neurons reacting to directional transitions from high to low spatial frequency'. 'Spatial frequency' is not a feature humans consciously use to categorize objects, but it is a feature that humans can make sense of. By contrast, if the features of the EXEMPLAR's referent are unrecognizable or too foreign to comprehend, it seems impossible to acquire the ability to judge new cases. This boils down to the question of whether modern image classifiers can be rendered understandable at all. I discuss this aspect in Sect. 5.

The previous discussion shows that the five desirable properties of EXEMPLARS can be mapped to properties of existing explanations. This suggests that EXEMPLARS are not only epistemically desirable, but may also be technically feasible. Ideally, explainability approaches would be able to generate EXEMPLARS with a clear scope of application that exemplify any selection of contextually relevant, intelligible features while disregarding irrelevant ones. However, it is unlikely that XAI will ultimately produce tailored EXEMPLARS for every conceivable explainability context. Still, it might be worth the effort in certain cases. And even as a purely theoretical construct, EXEMPLARS may serve as a useful foil for evaluating existing XAI methods and inspiring new approaches.

5 Are EXEMPLARS achievable?

In this section, I refute three potential objections to my proposal. They all cast doubt on whether EXEMPLARS are feasible in practice. However, even those readers who are not convinced by my responses to these objections may concede the following: Explanations provided by XAI approaches that are directed at image classifiers exemplify certain features they share with their referent. Analyzing what they exemplify can help to assess their suitability for producing understanding.

5.1 Objection 1: Current image classifiers exploit features of images that are too foreign to be intelligible to humans

The first objection is that EXEMPLARS cannot be created because current image classifiers, mainly those relying on neural network architectures, exploit features of images that are too foreign to be intelligible to humans. The discovery of so-called *adversarial examples* supports this suspicion. In the context of image classification, adversarial examples are images that are intentionally designed or chosen to lead to a wrong classification. Typically, an adversarial example is obtained by modifying a natural image that is classified correctly with a high probability. After the subtle modification, the image often appears unchanged to a human observer, but receives a different, apparently random classification from the model (Szegedy et al. 2014). Several other ways to obtain adversarial examples exist (e.g., Brown et al. 2017; Eykholt et al. 2018; Hendrycks et al. 2021; Nguyen et al. 2015; Sharif et al. 2016). They all seem to show that the behavior of image classifiers is too alien to be understood by humans.

However, some research suggests that the circumstances influence how humans perceive adversarials. For instance, some adversarial instances fool not only classifiers, but also time-constrained humans (Elsayed et al. 2018). In a study by Zhou and Firestone (2019), human subjects correctly predicted the model's classification of adversarial examples when they had to choose from a restricted number of categories. Furthermore, research on *network dissection* (Bau et al. 2017) suggests that the representations of deep neural networks at least sometimes correspond to human concepts. Bau et al. (2020) identified units in DCNN and Generative Adversarial Networks that correspond to familiar concepts such as 'snow', 'house', or 'oven'. However, not all units of deep neural networks can be matched to human concepts, the same unit might detect several concepts, and one concept can be learned by multiple (combinations of) units (Molnar 2022, ch. 9.1.2). Still, I take it that it is too early to conclude that modern image classifiers are hopelessly opaque. If it turns out that they respond to intelligible features, these features can be exemplified by EXEMPLARS. If not, this would not only make EXEMPLARS unachievable, but probably any other explanation of image classification as well.

5.2 Objection 2: Current image classifiers exploit abstract features that cannot be exemplified

The second objection is that EXEMPLARS cannot be created because current image classifiers, mainly those relying on neural network architectures, exploit abstract features of images that cannot be exemplified by EXEMPLARS. With respect to DCNN, Buckner (2018) argues that these models learn an abstract representation of each class. This abstract representation enables them to bracket so-called *nuisance factors* present in natural images. Nuisance factors are 'repeatable and systematic sources of variation that are not diagnostic of decision success' (Buckner 2019, p. 9). In image classification, such nuisances can take the form of variation in, e.g., 'size, pose, location, and rotation' (Buckner 2019, p. 9). This suggests that what hinders epistemic access to a classifier's representation is not the vast *quantity* of features the model exploits, but their *quality* of being abstract. Put another way, the difficulty does not arise because a classifier's representation of the 'robin' class incorporates an overwhelming amount of robin sizes, postures, viewing angles, and so on. Rather, it arises because the representation omits size, posture, or viewing angle altogether. At first glance, it seems impossible to exemplify features associated with such a representation. What would an example of the 'robin' class look like that shows a bird without any particular size or spatial orientation?

In answering this objection, it is helpful to distinguish between the complex workings of neural networks and the epistemic tools that help to understand them. Buckner (2018) is concerned with the former. In this context, there is something to the above objection. Abstract image features cannot be instantiated in isolation, but only in combination with the nuisance factors mentioned above. If the internal representation of a classifier is abstract, any attempt to visualize it will yield an inaccurate representation. However, this paper deals with tools to make image classifiers understandable. Visual examples can be valuable here. Granted, abstract features alone cannot be instantiated. Fortunately, they can still be exemplified. This is because nuisances can be disregarded like any other inert feature that the example instantiates. For instance, to exemplify the abstract feature of 'having a beak' (or pixels corresponding to it) can involve to also instantiate a particular beak position (cf. Elgin 2017, pp. 194–195). Such inert particular features can then be disregarded by appropriate stage setting. This is nothing unusual, even the most honored ornithologist will not (and cannot) use an abstract robin representation to explain the characteristic features of this species. Thus, I take it that there is no principled reason why abstract features associated with a target class cannot be exemplified by an EXEMPLAR. On the contrary, it may be difficult to gain epistemic access to such features without concretizing them through an example.

5.3 Objection 3: Achieving understanding via exemplification presupposes understanding

The third objection is that generating or choosing EXEMPLARS already presupposes the understanding that they shall produce (cf. Buckner 2018, p. 5344). An ornithologist can point to specific features of robins (and downplay others) only because she has a previous understanding of characteristic robin features. The same seems to hold for EXEMPLARS: Generating or selecting them requires access to contextually relevant features of their referent to decide what to exemplify, and what to disregard.

To answer this objection, one must distinguish between examples that serve as pedagogical tools to convey an understanding that has already been gained by someone else, and those that serve as a source for an understanding that was not previously available. In the latter case, explainability methods are used by experts to achieve a better understanding of how image classifiers work. Research by Cammarata et al. (2020) is a prime example. The authors analyzed each of the 10.000 neurons of a DCNN to gain a thorough understanding of how it processes image data. Among other things, they performed *feature visualization* as described in Sect. 3.3. For each neuron, the researchers visualized the features it maximally reacts to (Olah et al. 2017). In other words, they created an example that served as a source of an understanding which they were lacking before. This role of examples is also acknowledged by Elgin (2017, p. 190). Of course, to serve as a source of understanding, the example needs to adequately represent its referent. This can be ensured by relying on suitable procedures and by drawing on previous knowledge (cf. Elgin 2017, p. 190). If the XAI approaches that are used to produce EXEMPLARS work well, EXEMPLARS can be created without presupposing the understanding they provide. Importantly, this is not specific to EXEMPLARS. Also gaining understanding from the explanations of existing XAI approaches requires correct interpretation.

If the understanding in question was already achieved, EXEMPLARS can be used to convey it to others, including laypersons without much previous understanding. XAI contexts involve a heterogeneous group of stakeholders with varying degrees of background knowledge (Langer et al. 2021; Tomsett et al. 2018). That is, EXEMPLARS which convey understanding that experts already have, and which are tailored towards the needs of those different stakeholders, are a valuable tool for XAI.

6 Conclusion and outlook

XAI approaches that are directed at image classifiers aim to produce understanding by means of visual explanations. I have argued that analyzing what those explanations exemplify can help to assess their suitability for producing understanding. Furthermore, I suggested that such XAI methods should strive to produce what I call EXEMPLARS, i.e., examples that are tailored to context and mitigate the risk of misinterpretation.

However, there remain open questions. Obviously, research in computer science is required to explore whether and how EXEMPLARS can be created for different epistemic ends in image classification contexts. This includes to determine how different methods can be combined to create EXEMPLARS that are tailored to different contexts and accompanied by sufficient stage setting. More philosophical work is needed as well. While I have focused on one particular epistemic end, it will be fruitful to consider further epistemic ends when analyzing XAI approaches through the lens of exemplification. This can also help to specify what kinds of

EXEMPLARS would be suitable to meet those ends. Furthermore, more needs to be said about how context influences the suitability of EXEMPLARS. What is the impact of contextual factors on the features that an EXEMPLAR needs to exemplify? Do different stakeholders need different EXEMPLARS to achieve an epistemic end? Finally, further connections can be drawn between exemplification and understanding. I suggested to analyze the abilities an example affords. However, there are further aspects of understanding that may be facilitated by examples. As indicated in Sect. 2.2, it is widely held that understanding comes in degrees. Elgin (2017, pp. 58–59) distinguishes four dimensions along which the degree of understanding can vary: depth, breadth, facticity, and the acknowledgment of the relative significance of the facts within a body of information. Mapping those dimensions to the features that different explanations exemplify could reveal the degree of understanding they can provide.[8] For instance, saliency maps may be able to provide a deeper understanding than prototype-based explanations, and allow to assess the relative significance of different features to a classification. However, due to their limited scope, the understanding they provide may not be particularly broad.

In conclusion, exemplification has been largely overlooked in previous work on XAI. As I hope to have shown, it provides an epistemological framework that can shed light on various issues related to explainability and understanding, and deserves more attention from philosophers, computer scientists, and other scholars working in the field.

Acknowledgements I thank all members of the research project Explainable Intelligent Systems (EIS) for their helpful input. Special thanks go to Eva Schmidt for detailed feedback and encouragement, Barnaby Crook for pointing out several explainability methods of interest to my project, and two anonymous reviewers whose input significantly strengthened the paper. Finally, I thank audiences at the workshop 'Issues in XAI #5: Understanding Black Boxes–Interdisciplinary Perspectives' (Dortmund), at the workshop "Thought in Humans and Machines: Multidisciplinary Perspectives" (Hamburg), at the workshop 'Intelligent Systems in Context' (Bayreuth), at the 'First Luxembourg Workshop on Epistemology and Artificial Intelligence' (Luxembourg), at the reading group and the colloquium for theoretical philosophy (Dortmund), and at the EIS colloquium (Bayreuth/Dortmund/Saarbrücken) for the fruitful discussions.

Funding Open Access funding enabled and organized by Projekt DEAL. Work on this paper was funded by the Volkswagen Foundation grant AZ 98510 Explainable Intelligent Systems (EIS,https://explainable-intelligent.systems).

Data availability No datasets were generated or analyzed in this paper.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

References

Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Proceedings of 2018 advances in neural information processing systems (NeurIPS), vol 31. Curran Associates Inc, pp 1–11

Alqaraawi A, Schuessler M, Weiß P, Costanza E, Berthouze N (2020) Evaluating saliency map explanations for convolutional neural networks: a user study. in proceedings of the 25th international conference on intelligent user interfaces, IUI '20, New York, NY, USA. Association for Computing Machinery, pp 275–285. https://doi.org/10.1145/3377325.3377519

Alvarez-Melis D, Jaakkola TS (2018) On the robustness of interpretability methods. arXiv:1806.08049 [cs.LG]

Arrieta AB, Díaz-Rodríguez N, Ser JD, Bennetot A, Tabik S, Barbado A, Garcia S, Gil-Lopez S, Molina D, Benjamins R, Chatila R, Herrera F (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bau D, Zhu JY, Strobelt H, Lapedriza A, Zhou B, Torralba A (2020) Understanding the role of individual units in a deep neural network. Proc Natl Acad Sci 117(48):30071–30078. https://doi.org/10.1073/pnas.1907375117

Baumberger C (2019) Explicating objectual understanding: taking degrees seriously. J Gen Philos Sci 50(3):367–388. https://doi.org/10.1007/s10838-019-09474-6

Baumberger C, Beisbart C, Brun G (2017) What is understanding? An overview of recent debates in epistemology and philosophy of science. In: Grimm S, Baumberger C, Ammon S (eds) Explaining understanding: new perspectives from epistemology and philosophy of science. Routledge-Taylor & Francis, New York, Oxon, pp 1–34

Bau D, Zhou B, Khosla A, Oliva A, Torralba A (2017) Network dissection: quantifying interpretability of deep visual representations. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 3319–3327. https://doi.org/10.1109/CVPR.2017.354

Beisbart C (2021) Opacity thought through: on the intransparency of computer simulations. Synthese 199(3–4):11643–11666. https://doi.org/10.1007/s11229-021-03305-2

Beisbart C, Räz T (2022) Philosophy of science at sea: clarifying the interpretability of machine learning. Philos Compass 17(6):e12830. https://doi.org/10.1111/phc3.12830

Belle V, Papantonis I (2021) Principles and practice of explainable machine learning. Front Big Data 4:688969. https://doi.org/10.3389/fdata.2021.688969

Bien J, Tibshirani R (2011) Prototype selection for interpretable classification. Ann Appl Stat 5(4):2403–2424. https://doi.org/10.1214/11-aoas495

[8] I would like to thank an anonymous reviewer for pointing this out.

Boge FJ (2021) Two dimensions of opacity and the deep learning predicament. Minds Mach 32(1):43–75. https://doi.org/10.1007/s11023-021-09569-4

Brown TB, Mané D, Roy A, Abadi M, Gilmer J (2017) Adversarial patch. arXiv:1712.09665 [cs.CV]

Buckner CJ (2018) Empiricism without magic: transformational abstraction in deep convolutional neural networks. Synthese 195(12):5339–5372. https://doi.org/10.1007/s11229-018-01949-1

Buckner CJ (2019) Deep learning: a philosophical introduction. Philos Compass 14(10):e12625. https://doi.org/10.1111/phc3.12625

Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data Soc 3(1):1–12. https://doi.org/10.1177/2053951715622512

Cammarata N, Carter S, Goh G, Olah C, Petrov M, Schubert L (2020) Thread: circuits. Distill 5(3). https://doi.org/10.23915/distill.00024

de Regt HW (2015) Scientific understanding: truth or dare? Synthese 192(12):3781–3797. https://doi.org/10.1007/s11229-014-0538-7

Dhurandhar A, Chen PY, Luss R, Tu CC, Ting P, Shanmugam K, Das P (2018) Explanations based on the missing: towards contrastive explanations with pertinent negatives. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Proceedings of 2018 advances in neural information processing systems (NeurIPS), vol 31. Curran Associates, Inc, pp 1–12

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani , Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2021) An image is worth 16x16 words: transformers for image recognition at scale. pp 1–22. arXiv:2010.11929 [cs.CV]

Elgin CZ (2017) True enough. MIT Press, Cambridge

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: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Proceedings of 2018 advances in neural information processing systems (NeurIPS), vol 31. Curran Associates, Inc, pp 1–11

Erhan D, Bengio Y, Courville A, Vincent P (2009) Visualizing higher-layer features of a deep network. Technical report no. 1341, University of Montreal

Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C, Prakash A, Kohno T, Song D (2018) Robust physical-world attacks on deep learning visual classification. In: Proceedings of the 2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 1625–1634. https://doi.org/10.1109/CVPR.2018.00175

Fleisher W (2022) Understanding, idealization, and explainable AI. Episteme 19(4):534–560. https://doi.org/10.1017/epi.2022.39

Fong RC, Vedaldi A (2017) Interpretable explanations of black boxes by meaningful perturbation. In: Proceedings of the 2017 IEEE international conference on computer vision (ICCV). pp 3449–3457. https://doi.org/10.1109/iccv.2017.371

Fujiyoshi H, Hirakawa T, Yamashita T (2019) Deep learning-based image recognition for autonomous driving. IATSS Res 43(4):244–252. https://doi.org/10.1016/j.iatssr.2019.11.008

Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, Wichmann FA (2020) Shortcut learning in deep neural networks. Nat Mach Intell 2(11):665–673. https://doi.org/10.1038/s42256-020-00257-z

Ghorbani A, Abid A, Zou J (2019) Interpretation of neural networks is fragile. Proc AAAI Conf Artif Intell 33(1):3681–3688. https://doi.org/10.1609/aaai.v33i01.33013681

Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press, Cambridge

Goodman B, Flaxman S (2017) European union regulations on algorithmic decision-making and a 'right to explanation'. AI Mag 38(3):50–57. https://doi.org/10.1609/aimag.v38i3.2741

Goyal M, Knackstedt T, Yan S, Hassanpour S (2020) Artificial intelligence-based image classification methods for diagnosis of skin cancer: challenges and opportunities. Comput Biol Med 127:104065. https://doi.org/10.1016/j.compbiomed.2020.104065

Grimm SR (2011) Understanding. In: Bernecker S, Pritchard D (eds) Routledge companion to epistemology. Routledge, New York, pp 84–94

Gu J, Tresp V (2019) Saliency methods for explaining adversarial attacks. In: Human-centric machine learning, NeurIPS 2019 workshop. arXiv:1908.08413 [cs.CV]

Hendrycks D, Zhao K, Basart S, Steinhardt J, Song D (2021) Natural adversarial examples. In: Proceedings of the 2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 15262–15271. https://doi.org/10.1109/CVPR46437.2021.01501

Hills A (2016) Understanding why. Noûs 50(4):661–688. https://doi.org/10.1111/nous.12092

Khalifa K (2017) Understanding, explanation, and scientific knowledge. Cambridge University Press, Cambridge

Kim B, Khanna R, Koyejo O (2016) Examples are not enough, learn to criticize! criticism for interpretability. In: Proceedings of the 30th international conference on neural information processing systems, NIPS'16, Red Hook, NY, USA. Curran Associates Inc, pp 2288–2296

Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, Sayres R (2018) Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, vol 80 of proceedings of machine learning research, Stockholmsmässan, Stockholm, Sweden. PMLR. pp 2668–2677

Koh PW, Liang P (2017) Understanding black-box predictions via influence functions. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, vol 70 of proceedings of machine learning research. PMLR. pp 1885–1894

Krishnan S, Wu E (2017) PALM: machine learning explanations for iterative debugging. In: Proceedings of the 2nd workshop on human-in-the-loop data analytics, HILDA '17, New York, NY, USA. Association for Computing Machinery (ACM), pp 1–6. https://doi.org/10.1145/3077257.3077271

Kvanvig JL (2003) The value of knowledge and the pursuit of understanding. Cambridge studies in philosophy. Cambridge University Press, Cambridge

Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, Sesing A, Baum K (2021) What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artif Intell 296:1–24. https://doi.org/10.1016/j.artint.2021.103473

Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR (2019) Unmasking clever Hans predictors and assessing what machines really learn. Nat Commun 10(1):1096. https://doi.org/10.1038/s41467-019-08987-4

Law J, Lynch M (1988) Lists, field guides, and the descriptive organization of seeing: birdwatching as an exemplary observational activity. Hum Stud 11(2–3):271–303. https://doi.org/10.1007/bf00177306

LeCun Y, Bengio Y (1995) Convolutional networks for images, speech, and time series. In: Arbib MA (ed) The handbook of brain theory and neural networks. MIT Press, Cambridge, pp 255–258

Lipton ZC (2018) The mythos of model interpretability. Queue 16(3):31–57. https://doi.org/10.1145/3236386.3241340

Mann S, Crook B, Kästner L, Schomäcker A, Speith T (2023) Sources of opacity in computer systems: towards a comprehensive taxonomy. In: 2023 IEEE 31st international requirements engineering conference workshops (REW), Hannover. IEEE. pp 337–342. https://doi.org/10.1109/REW57809.2023.00063

McDermid JA, Jia Y, Porter Z, Habli I (2021) Artificial intelligence explainability: the technical and ethical dimensions. Philos Trans R Soc A Math Phys Eng Sci 379(2207):20200363. https://doi.org/10.1098/rsta.2020.0363

Miller T (2019) Explanation in artificial intelligence: insights from the social sciences. Artif Intell 267:1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc 3(2):1–21. https://doi.org/10.1177/2053951716679679

Molnar C (2022) Interpretable machine learning. A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/

Mordvintsev A, Olah C, Tyka M (2015) Inceptionism: going deeper into neural networks. https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Mueller ST (2020) Cognitive anthropomorphism of AI: how humans and computers classify images. Ergonom Des Q Hum Fact Appl 28(3):12–19. https://doi.org/10.1177/1064804620920870

Newman M (2017) An evidentialist account of explanatory understanding. In: Grimm SR, Baumberger C, Ammon S (eds) Explaining understanding: new perspectives from epistemology and philosophy of science. Taylor & Francis, New York, pp 190–211

Nguyen A, Yosinski J, Clune J (2015) Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 427–436

Nguyen A, Dosovitskiy A, Yosinski J, Brox T, Clune J (2016) Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Lee D, Sugiyama M, Luxburg U, Guyon I, Garnett R (eds) Proceedings of 2016 advances in neural information processing systems (NeurIPS), vol 29. Curran Associates, Inc, pp 1–9

Nguyen A, Clune J, Bengio Y, Dosovitskiy A, Yosinski J (2017) Plug & play generative networks: conditional iterative generation of images in latent space. In: Proceedings of the 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp 3510–3520. https://doi.org/10.1109/CVPR.2017.374

Nyrup R, Robinson D (2022) Explanatory pragmatism: a context-sensitive framework for explainable medical AI. Ethics Inf Technol. https://doi.org/10.1007/s10676-022-09632-3

Olah C, Mordvintsev A, Schubert L (2017) Feature visualization. Distill 2(11). https://doi.org/10.23915/distill.00007

Páez A (2019) The pragmatic turn in explainable artificial intelligence (XAI). Minds Mach 29(3):441–459. https://doi.org/10.1007/s11023-019-09502-w

Peterson RT (1980) A field guide to the birds: eastern and central North America. Houghton Mifflin Harcourt, Boston

Petsiuk V, Jain R, Manjunatha V, Morariu VI, Mehra A, Ordonez V, Saenko K (2021) Black-box explanation of object detectors via saliency maps. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). pp 11443–11452. https://doi.org/10.1109/CVPR46437.2021.01128

Rabold J, Deininger H, Siebers M, Schmid U (2020) Enriching visual with verbal explanations for relational concepts—combining LIME with aleph. In: Cellier P, Driessens K (eds) Machine learning and knowledge discovery in databases. Springer International Publishing, Cham, pp 180–192. https://doi.org/10.1007/978-3-030-43823-4_16

Ribeiro MT, 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, pp 1135–1144. https://doi.org/10.1145/2939672.2939778

Ribeiro MT, Singh S, Guestrin C (2018) Anchors: high-precision model-agnostic explanations. In: Proceedings of the thirty-second AAAI conference on artificial intelligence, pp 1527–1535. https://doi.org/10.1609/aaai.v32i1.11491

Riggs WD (2003) Understanding 'virtue' and the virtue of understanding. Intellectual virtue. Oxford University Press. pp 203–226. https://doi.org/10.1093/acprof:oso/9780199252732.003.0010

Schubert L, Voss C, Cammarata N, Goh G, Olah C (2021) High-low frequency detectors. Distill. https://doi.org/10.23915/distill.00024.005

Schwalbe G, Finzel B (2023) A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. Data Min Knowl Discov. https://doi.org/10.1007/s10618-022-00867-8

Sharif M, Bhagavatula S, Bauer L, Reiter MK (2016) Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp 1528–1540. https://doi.org/10.1145/2976749.2978392

Simonyan K, Vedaldi A, Zisserman A (2014) Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv:arXiv:1312.6034v2 [cs.CV]

Smolensky P (1988) On the proper treatment of connectionism. Behav Brain Sci 11(1):1–23. https://doi.org/10.1017/s0140525x00052432

Speith T (2022) A review of taxonomies of explainable artificial intelligence (XAI) methods. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, pp 2239–2250. https://doi.org/10.1145/3531146.3534639

Strevens M (2013) No understanding without explanation. Stud Hist Philos Sci Part A 44(3):510–515. https://doi.org/10.1016/j.shpsa.2012.12.005

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks. arXiv:1312.6199 [cs.CV]

Szeliski R (2022) Computer vision. Algorithms and applications, 2nd edn. Springer, Cham

Tomsett R, Braines D, Harborne D, Preece A, Chakraborty S (2018). Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. In: 2018 ICML workshop on human interpretability in machine learning (WHI 2018), Stockholm, pp 8–14. https://doi.org/10.48550/arXiv.1806.07552

Wen D, Khan SM, Xu AJ, Ibrahim H, Smith L, Caballero J, Zepeda L, de Blas Perez C, Denniston AK, Liu X, Matin RN (2021) Characteristics of publicly available skin cancer image datasets: a systematic review. Lancet Digit Health 4(1):e64–e74. https://doi.org/10.1016/s2589-7500(21)00252-1

White A, Ngan KH, Phelan J, Afgeh SS, Ryan K, Reyes-Aldasoro CC, d'Avila Garcez A (2021) Contrastive counterfactual visual explanations with overdetermination. arXiv:2106.14556 [cs.CV]

Wilking R, Jakobs M, Morik K (2022) Fooling perturbation-based explainability methods. In: Workshop on trustworthy artificial intelligence as a part of the ECML/PKDD 22 program, Grenoble, France, IRT SystemX [IRT SystemX], pp 1–16

Woodward J (2004) Making things happen: a theory of causal explanation. Oxford University Press, Oxford

Wu X, Zhang X (2016) Responses to critiques on machine learning of criminality perceptions (addendum of arxiv:1611.04135). arXiv:1611.04135 [cs.CV]

Yadav SS, Jadhav SM (2019) Deep convolutional neural network based medical image classification for disease diagnosis. J Big Data 6(1):1–18. https://doi.org/10.1186/s40537-019-0276-2

Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding neural networks through deep visualization. arXiv:1506.06579 [cs.CV]

Zednik C (2021) Solving the black box problem: a normative framework for explainable artificial intelligence. Philos Technol 34(2):265–288. https://doi.org/10.1007/s13347-019-00382-7

Zhou Z, Firestone C (2019) Humans can decipher adversarial images. Nat Commun. https://doi.org/10.1038/s41467-019-08931-6

Zhou J, Chen F, Holzinger A (2022) Towards explainability for AI fairness, xxAI—beyond explainable AI. Springer, Cham, pp 375–386. https://doi.org/10.1007/978-3-031-04083-2_18

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.