Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-03277-x OPEN

What do algorithms explain? The issue of the goals and capabilities of Explainable Artificial Intelligence (XAI)

Moritz Renftle[1], Holger Trittenbach[2], Michael Poznic[3]✉ & Reinhard Heil[3]

The increasing ubiquity of machine learning (ML) motivates research on algorithms to "explain" models and their predictions—so-called Explainable Artificial Intelligence (XAI). Despite many publications and discussions, the goals and capabilities of such algorithms are far from being well understood. We argue that this is because of a problematic reasoning scheme in the literature: Such algorithms are said to complement machine learning models with desired capabilities, such as *interpretability* or *explainability*. These capabilities are in turn assumed to contribute to a goal, such as *trust* in a system. But most capabilities lack precise definitions and their relationship to such goals is far from obvious. The result is a reasoning scheme that obfuscates research results and leaves an important question unanswered: What can one expect from XAI algorithms? In this paper, we clarify the modest capabilities of these algorithms from a concrete perspective: that of their users. We show that current algorithms can only answer user questions that can be traced back to the question: "How can one represent an ML model as a simple function that uses interpreted attributes?". Answering this core question can be trivial, difficult or even impossible, depending on the application. The result of the paper is the identification of two key challenges for XAI research: the approximation and the translation of ML models.

[1] Independent Scholar, Karlsruhe, Germany. [2] Neurocat GmbH, Rudower Chaussee 29, 12489 Berlin, Germany. [3] Karlsruhe Institute of Technology (KIT), Institute for Technology Assessment and Systems Analysis (ITAS), PO Box 3640, 76021 Karlsruhe, Germany. ✉email: michael.poznic@kit.edu

Introduction

Achieving high predictive accuracy has been the lynchpin of Machine Learning (ML) research and a fundamental requirement for using ML in practice. However, the increasing pervasiveness of ML has led to concerns among policymakers, researchers, and society about the ethical and legal implications of using ML without an *understanding* of the behavior and limitations of the models generated (cf. Goodman and Flaxman, 2017). This has fostered research, so-called Explainable Artificial Intelligence (XAI), on algorithms that generate "explanations" of ML. The explanations in turn are expected to improve ML users' and researchers' understanding.[1]

Over time, XAI algorithms have turned into a supposed panacea for virtually every concern that may be raised in the context of ML. There is a common scheme of reasoning regarding XAI algorithms. First, one points out an important capability that an ML model does *not* possess. Typical capabilities in focus are *interpretability*, *explainability*, *transparency*, or *comprehensibility* (cf. Páez, 2019; Szczepański et al., 2021). The argument then goes that XAI algorithms supplement the respective capability through some form of explanation. This explanation in turn is expected to help in achieving a postulated *goal*, such as building trust in a model (Arrieta et al., 2020) or delivering reasons for using the model in a specific context (Adadi and Berrada, 2018).

This scheme is problematic for various reasons, and we address both the scheme and some of its problems in the section "The reasoning scheme". However, what seems to be an even more pressing problem is that there are diverse contextual meanings of the terms used to describe the algorithms' capabilities. For example, there is no consensus on what "interpretability" means in the context of ML, and which criteria an XAI method must fulfill in order to make an ML model interpretable; the same holds for most of the remaining capabilities. This issue has been highlighted in the scientific literature ever since (cf. Lipton, 2018; Krishnan, 2020; Robbins, 2019; Erasmus et al., 2021). If one is to depart from the problematic reasoning scheme, the question remains what one can reasonably expect from XAI algorithms. In this paper, we seek an answer to this question through an interdisciplinary perspective from computer science, philosophy, and technology assessment. We analyze two modest capabilities that play an important role in our discussions: (1) the algorithms' input attributes have to be *interpreted*, and (2) the requirement that functions which translate attributes into interpreted ones have to be *simple*.[2]

We find that the modest capabilities of actual interpretation and simplicity can be delivered by XAI algorithms, and so they indeed help to understand what ML models can do and are in fact doing in the context of specific applications. These properties are more modest capabilities than the discussed capabilities in the literature, and we submit that it is a more realistic aim for XAI research to focus on these capabilities.

There is a recent trend to look more closely at the users of XAI algorithms to address the issue of XAI algorithms' goals and capabilities. Tomsett et al. (2018) and Zednik (2021) focus on different stakeholders and the questions they might pose when confronted with ML applications. In line with these proposals, our approach is to consider the perspectives of users of ML and to engage in a thought experiment to develop our argument. The thought experiment in which we invite the reader to participate in the practical situation of using an email spam filter.

We use this thought experiment as a first step in clarifying two important capabilities, actual interpretation and simplicity, that are better suited to answering the quest for XAI algorithms' capabilities. More precisely, our aim is to inquire which capabilities one can reasonably expect to be delivered by XAI algorithms.

We start with a general perspective of a curious human being confronted with a technical tool that produces remarkable results. Such a human user of ML models strives for understanding, meaning that they have questions about these models. We suggest interpreting XAI algorithms as methods that help to answer these questions, or at least some disambiguated versions of them. However, we show that algorithms can only answer a very specific type of question. From this viewpoint, clarifying the capabilities of XAI algorithms means, (i) to collect questions that the curious user might have about ML models, and (ii) to identify the subset of these questions that XAI algorithms help to answer. Further, clarifying the capabilities of existing XAI algorithms means, (iii) to examine what is difficult about the questions identified in (ii), i.e., to identify the challenges for these algorithms, and (iv) to assess how far the challenges are met by existing algorithms.

This paper is structured accordingly. We address the problems of the reasoning scheme in the section "The reasoning scheme". In the section "Questions about ML models", we introduce the thought experiment of the spam filter and discuss the questions that users might ask about ML models. In the section "Questions addressed by XAI algorithms", we find one of these questions to be the main question currently addressed by XAI algorithms. Answering this question reveals two general challenges, one of translation and one of approximation, which we present in the section "Challenges for XAI algorithms". In the section "The state of XAI algorithms", we review how far existing XAI algorithms meet these challenges. Section "Conclusions" concludes.

The reasoning scheme

As outlined in the Introduction, there is a common reasoning regarding XAI algorithms in the scholarly literature. An important *capability* of an ML model is claimed to be missing, such as *interpretability*, *explainability*, or *comprehensibility*. The algorithms should then supplement the capability of interest via an explanation, and this explanation finally fosters the achievement of a *goal*, such as building trust or delivering reasons for using the model.

One can find specific criticism of accounts of some of the capabilities attributed to XAI in the literature. One strategy of critique is to challenge the definition of the term in question (cf. Krishnan, 2020) by pointing out some weaknesses and proposing an alternative. A further strategy is to differentiate the terms into different senses (cf. Lipton, 2018). We use an alternative strategy. We criticize the reasoning scheme that connects capabilities to goals, and we discuss realistic expectations of what XAI can achieve, based on the diagnosis that most proposed capabilities are too demanding.

To make our criticism concise, we focus on *explainability* in this section. This is an arbitrary choice, but we expect our arguments to hold for other capabilities as well. One can find several discussions of capabilities in the literature; examples are discussions of interpretability (Lipton, 2018; Krishnan, 2020, Erasmus et al., 2021; Fleisher, 2022), explainability (Arrieta et al., 2020; Fleisher, 2022), and of related notions such as explicability (Robbins, 2019) and transparency (Fleisher, 2022). In the reconstruction of the reasoning scheme below, the term 'explainable' can easily be exchanged for 'explicable', 'interpretable', etc.[3]

The reasoning scheme that we reconstruct in the following paragraphs is not explicitly proposed in the literature. It is rather implicit in the argumentation of many publications. For the goal of trust or trustworthiness, one can point to several publications. We mention only a few here. In Ribeiro et al. (2016), "trust" is used in the title. Furthermore, the authors state:

> We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted. (Ribeiro et al., 2016, p. 1135).

In another publication, one finds this instance of 'trust':

> Explanations of machine learning models and predictions can serve many functions and audiences. Explanations can […] verify and improve the functionality of a system […] and enhance the trust between individuals subject to a decision and the system itself. (Mittelstadt et al., 2019, p. 280).

In some places, explainability is even declared as a necessary prerequisite for humans to trust ML models.

> In order for humans to trust black-box methods, we need explainability—models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions. (Gilpin et al., 2018).

Molnar (2020) is a frequently referenced introductory book to interpretable ML, and here one finds:

> If you can ensure that the machine learning model can explain decisions, you can also check the following traits more easily: Fairness […] Privacy […] Reliability […] Causality […] Trust: It is easier for humans to trust a system that explains its decisions compared to a black box.

In Erasmus et al. (2021), the authors remain somewhat uncommitted to the requirement of generating trust by fostering the capability of ML models, but the term is mentioned in their conclusion.

> Conceiving of explanation according to those accounts offered within the philosophy of science disentangles the accuracy-explainability trade-off problem in AI, and in doing so, deflates the apparently paradoxical relationship between trust and accuracy which seems to plague debates in medical AI and the ML literature. *If* it is simply that explainability is required for trust, there is no cause for worry, since highly accurate (and potentially complex) MAIS [medical AI systems] are just as explainable as simple ones. (Erasmus et al., 2021, p. 857).

From a philosophical perspective, one can reconstruct the common reasoning as follows:

i) Users of an ML model can give an explanation for the predictions of the model if and only if the model is explainable for those users.

ii) The goal of such an explanation is to generate trust in actors who are affected by the decisions made based on the model predictions.

iii) Users can generate trust in actors if and only if the actors accept the explanation as justification for the decisions made based on the model prediction.

iv) If the actors accept the explanation as a justification for the decisions made based on the model prediction, then the explanation generates trust in the actors.

In the following, we question whether such a capability is indeed required to achieve the goal. On the one hand, one can ask for further supporting reasons for (ii) as well as for (iii) or (iv). Especially, the assumed relation between explainability and trust is oftentimes not well defined. This becomes clear in a more stringent version of the reconstruction of the reasoning scheme:

(1) Users can give a justification for the decisions based on the predictions of an ML model if the model is explainable for its users.
(2) Other people—actors—trust in the model if these actors are convinced by the justification the users give.
(3) The justification works by "explaining" the model to the actors. Thereby the model becomes explainable also for the actors, and establishes trust among the actors.

In the second version, we fill the gap between capability and goal by an intermediate step of justification. However, there is no compelling reason why the justification is sufficient to convince the actors—even if the ML model is explainable for the users. Next, the capability also may not be necessary to achieve the goal of trust.[4] For instance, achieving trust in a technical system is feasible through other means than via the capability of the algorithm. One way that actors build trust could be through the reputation of those promoting a particular ML model, or supported by the past performances of either the same model or other models promoted by the same parties.

As a result of these considerations, we can state a short interim conclusion. The common reasoning scheme is insufficient to approach the issue of XAI goals and capabilities. First, one has to disentangle goals and capabilities. Second, the capabilities need to be scrutinized in detail. Third, the relation between capability and goals is not properly articulated, and it is by no means clear that the capability directly leads to the respective goal. Our focus here is to clarify how one can reach justified claims about the capabilities of XAI. Especially we focus on capabilities that help users and actors to understand ML models. We deem all of these clarifications necessary to inform discussion on adequate goals for XAI, however, we leave this discussion for another occasion.

Questions about ML models

There are discussions of user questions in the literature. For example, Liao et al. (2020) discuss the suitability of explanations as being question-dependent in the context of application-oriented research. Therefore, users can ask many questions about ML models. Here, we introduce seven questions that one may ask when confronted with a specific ML model. Even though we found these questions by philosophical methods rather than by using empirical methods, some of them appear frequently in the literature (e.g., Gunning, 2017; Hoffman et al., 2018). We do not claim that any user has these questions in mind. What we claim is that the questions are reasonable and not arbitrarily invented.

We reached these questions by engaging in a thought experiment of a use case of ML. In this thought experiment, a user of an email program is confronted with the issue of the detection of spam emails through an ML model. The questions later serve to delineate the capabilities of XAI algorithms. While they are based on one example, we deem them to be general and also applicable to other ML use cases.

Thought experiment of the spam filter. Assume a spam filter *S* is an ML model using the methodology of supervised ML. *S* classifies emails into two disjoint groups, *spam* and *no spam*. Further, there is a user, Alice, who sees an email *M* which *S* classifies as *spam*. Alice is a curious email recipient and has an interest in understanding the ML model. Given this small set of pieces of information, which questions are reasonable for Alice to ask? The central task of the model is to distinguish regular emails, which we call "normal" ones, from spam emails, and Alice is confronted with a new email *M*. So, Alice may reasonably ask only a small number of questions about issues related to these things. The list we provide is comprehensive. However, we do not claim that it is

exhaustive. We argue that the following questions are relevant for Alice:

Q1: Why is *M* classified as *spam* by *S*?
Q2: How does *S* distinguish *spam* from *no spam*?
Q3: What distinguishes *spam* from *no spam*?
Q4: How does *S* work?
Q5: Does *S* work like an alternative spam filter *S** Alice has used in the past?
Q6: Alice thinks that *M* is not spam. Why does *S*'s classification differ from Alice's opinion?
Q7: What distinguishes spam emails from normal emails?

Ambiguities of the questions. Intentionally, we have only made the assumption that Alice is curious, so far. In particular, she is one of the actors seeking a higher goal such as trust in the spam filter (see section "The reasoning scheme"), and we do not expect her to have special knowledge in computer science or philosophy. Naturally, to ask questions from such a layperson's perspective leads to formulations that may turn out to be ambiguous depending on the context in which they are asked. Let us illustrate this with an example.

A computer scientist, for instance, may differentiate Q1 between asking for an explanation ex-ante, i.e., before the model has returned a prediction of *M*, versus an ex-post explanation when the model prediction is known. In both cases the explanation is *local*, i.e., it is specific for *M*. However, in *an ex-ante explanation*, the prediction of *M* must not be part of the explanans. In particular, any XAI method that relies on the prediction of *M*, e.g., a counterfactual explanation (Verma et al., 2022), does not apply. Instead, an ex-ante explanation requires a partial answer of Q2, and then to derive the specific explanation for *M*. To illustrate, if an answer to Q2 is that *S* distinguishes *spam* from *no spam* by the email domain of the sender, then an ex-ante explanation for Q1 is that *M* has a specific domain D. Note that this ex-ante explanation requires only a partial answer of Q2; in particular, it does not require us to explain the prediction of *S* for any $M' \neq M$. On the other hand, one can deem Q1 to ask for *an ex-post explanation*. In this case, one may formulate an explanation based on the prediction of *S* on *M*. For instance, one may create a counterfactual that searches for minor modifications of *M* such that the prediction of S changes. An example of a counterfactual would be to alter a few words in the body of the message such that M is classified as *no spam*; the difference between the original and modified *M* is the explanans.

A philosopher may have a different disambiguation in mind. For instance, the type of explanation asked for can be of importance. For Q1, one could specifically ask for, say, a deductive nomological explanation: "What is a deductive nomological explanation for *S* classifying *M* as *spam*?". Here, one has a comprehensive theory of explanation at one's disposal to disambiguate the questions. The refined question may indeed be less ambiguous in the sense that it is more likely to get a philosophically satisfying answer.[5]

Q3 is special, since it is not necessarily related to ML. In one sense it is a general question similar to questions of the form "what is F?", where F is a variable for a general term that could be interpreted to be *spam* but also to be something more theoretical, such as *knowledge*, or something more practical like *the meaning of life*. Laypeople ask these kinds of questions, but also philosophers usually ask them. And since its beginning from Plato onwards, Western philosophy has addressed these apparently simple but difficult to answer questions. What interests us in the context of this paper is whether XAI can make any contribution to answering such a question.[6]

We deem both directions of disambiguation, for computer science and for philosophy, as indispensable in advancing the debate around XAI, and we also strive for disambiguation of some of the questions below. However, we do not attempt to account for all the various ways of disambiguation in this paper. We rather see the merits of discussing Q1-Q7 in approaching the debate around XAI differently and departing from a problematic reasoning scheme. We show that turning to a questions-centered approach can indeed bring clarity to the debate, and that it helps to reveal unrealistic and unproductive expectations.

We now turn to the questions Alice asks, and clarify what is required to answer them. In the end, we arrive at one question with an improved formulation that we deem answerable in light of the current literature on XAI: "How can one represent an ML model as a simple function that uses interpreted attributes?". It is a reformulation of Q2.

Discussing the questions. **Q1: Why is *M* classified as *spam* by *S*?**

To answer this question, it is necessary to clarify what exactly *S* is, and on what basis it makes classifications.

Mathematically, *S* is a function that maps from emails to either the label *spam* or *no spam*. Many such functions exist, but only some of them are accurate spam filters. To find one of these accurate functions, our example uses supervised ML.

Supervised ML is a general methodology for predicting a target variable *Y* based on an observable variable *X*. The goal is to find a function that maps values of the observable variable to values of the target variable, with low error. To find such a function, supervised ML requires *training data*. Training data is a set of pairwise observations (x_i, y_i) of *X* and *Y*, where *i* is an integer identifying the pair. In our case, the x_i represent emails, and the y_i are either spam or no spam. x_i can represent whole emails, e.g., as a sequence of zeros and ones in binary format. But x_i can also consist of only some attributes of an email, e.g., the domain of the sender or the number of words in the email body. Supervised ML assumes that the training data is representative of the joint distribution of the two variables *X* and *Y*. There are different notions of representativeness, but the most frequent is that, (a) the observations in the training data have been sampled independently and identically distributed (i.i.d.) from *X* and *Y*, and that (b) the number of observations is "large enough".

Based on the training data, supervised ML selects a function to map from *X* to *Y* with low prediction error. The function selection typically involves two steps: (i) choosing a set of candidate functions, and (ii) finding the optimal function within this set. In step (i), one chooses a *function type*, i.e., a set of functions whose equations have a similar form. Examples of function types are linear functions, decision tree functions, or neural network functions. For the chosen function type, one further chooses so-called hyperparameters, e.g., the number of coefficients of a linear function. The result of step (i) is a function with a number of free parameters, often called *an ML model*. In step (ii), one then uses an ML algorithm to optimize the remaining parameters of the model using the training data.

The algorithm returns *a trained ML model*, e.g., the model together with the parameters that lead to the lowest prediction error for the observations in the training data or a regularized, sparse model. If the training data is indeed representative of the joint distribution of *X* and *Y*, then this model will also have a low prediction error for new observations of *X* and *Y*.

In view of this, we can now answer why *S* makes a particular classification for M (Q1). The main point of the answer is related to the fact that the classifications of *S* are the result of a specific supervised ML methodology. This methodology involves many decisions and assumptions, e.g., regarding the collection of the

training data or the selection of candidate ML models. These decisions and assumptions determine both the overall prediction performance of *S* and the individual errors of *S*.

Q2: How does *S* distinguish *spam* from *no spam*?

This question can be answered by pointing to the attributes of *M* and the function by which *S* distinguishes between the labels *spam* and *no spam*.

An important distinction is between attributes that humans can interpret or communicate and attributes that refer only to the digital representation of a data item and exist only for the purpose of ML. We will call the former attributes *interpreted attributes* and the latter *technical attributes*. In the case of an email, interpreted attributes include the sender domain, letters, words, phrases, or the word count. A technical attribute would be, for example, the result of a principal component analysis in the form of numerical values (Abdi and Williams, 2010). Technical attributes can lack human interpretation; e.g., one might not be able to tell what the first principal component means. Another example for a technical attribute of emails is their representations as high-dimensional numerical vector embedding (Le and Mikolov, 2014).

One way to think about interpreted attributes is to see them related to competencies of users in mastering a language. In this case, the terms used to describe attributes have to be understood. This understanding is a different understanding compared to the understanding of the overall ML models. In particular, one has to understand a particular subset that is used in parts of the algorithm. This is a different form of understanding, namely a linguistic or symbolic understanding (cf. Baumberger et al., 2017). Apart from the understanding of the attributes the individual users have, a common understanding of the attributes is crucial so that users can communicate with each other using the terms used to describe the attributes.

Another aspect to consider with interpreted attributes is the context in which the questions are asked. There is a difference between a "system administrator" that maintains an email server and a "layperson". While IP addresses of the sender as email header would count as interpreted attributes for the system administrator, they might be technical attributes for the layperson.

There are cases where the difference between technical and interpreted attributes is less clear. Think of models that classify image data. In this case, is "an RGB value of a pixel in the image" an interpreted attribute? Are superpixels (Ribeiro et al., 2016), i.e., a group of connected pixels that are similar to each other, interpreted attributes? The answers to these questions depend on the context. In the medical imaging domain, a pathologist might understand superpixels since they have a specific interpretation. In another domain, e.g., autonomous driving, superpixels might count as technical attributes. There, interpreted attributes might rather be the illumination of an image or the presence of objects in specific locations of the image. For the purpose of discussing Q2, it suffices to assume that there is *some* context in which the user can understand the attributes to some degree. Then, a user can achieve what Khalifa (2017) coins as *generic understanding*.[7] Thus, attributes are *interpreted attributes* as long as there is some context where a user can understand them, and *technical attributes* otherwise.

One can answer Q2 in different ways. A simple answer is to write down the whole ML model *S*, as a complex mathematical expression that involves technical attributes of emails.

But this answer has two problems. First, *S* might use many more numeric parameters and attributes of emails than a human mind can comprehend. In this case, the mathematical definition of *S* is a correct answer to Q2, but not one that satisfies users such as Alice. Second, even if the mathematical definition of an ML model is "simple", it might still not satisfy users if it only refers to technical attributes and not to interpreted ones. The reason is that users of *S* typically have prior knowledge and expectations regarding the phenomenon "spam" to which they want to compare *S*. This requires a definition of *S* that ties to users' prior knowledge, i.e., that uses *interpreted* attributes. For example, if the function of *S* relies only on the first principal component of an email vector embedding, it remains unclear whether this value is correlated with some interpreted attribute of emails. Overall, a good answer to Q2 that can be offered to users would be a simple function that uses interpreted attributes of emails but still accurately describes how the spam filter *S* classifies emails. Note that such a function need not exist, or that there may be several functions that describe the original ML model with similar accuracy.

Q3: What distinguishes *spam* from *no spam*?

So far, we have focused on *S* and classification. Q3 now changes the perspective. As mentioned above, the answer to what discerns *spam* from *no spam* is independent of *S* and also independent of any other ML model. One could approach this question from a purely philosophical perspective and engage in a conceptual analysis of the notion of spam. Were one to conclude such an analysis then the distinction between *spam* and *no spam* might be made with the help of conceptual methods only. As our paper is based on an interdisciplinary approach that combines philosophy and computer science, we continue with an alternative proposal that emphasizes computer science methods. Let us assume that it depends on the training data for *S* of what *spam* is.

Then, a potential clarification of Q3 is: "Which functions exist that distinguish the emails labeled *spam* from those labeled *no spam* within the training data?". As discussed before, users require functions that are simple and use interpreted attributes. Hence, we only consider such functions in the following discussion.

The complete answer to Q3 is to specify *all* simple functions that use interpreted attributes and accurately distinguish *spam* from *no spam* within the training data. Typically, the number of such functions may be very large or even infinite, depending on the type and number of attributes in the training data. Thus, examining all these functions is infeasible for users.

So, a less accurate but more useful answer to Q3 is a small, diverse set of functions that are simple, use interpreted attributes and accurately distinguish *spam* from *no spam* within the training data. There are two ways to generate such a set. First, one can create new training data that only uses interpreted attributes of emails. One then runs an ML algorithm that optimizes a simple function for distinguishing *spam* from *no spam* based on these attributes. One repeats the process with other ML algorithms until one has a diverse set of functions. Second, one can also start with a complex ML model, like *S*, and then derive different simple functions from it that use interpreted attributes. This is equivalent to answering Q2. One can further diversify the results by deriving the functions from different complex ML models.

Q4: How does *S* work?

This question asks for the mechanics of *S*. A simple answer is that *S* is a function that maps from emails to labels. But one can also interpret this question as asking for the mechanism of how to derive *S* from data. With this interpretation, the answer to Q4 includes the answer to Q1, i.e., the selection of a predictive function based on training data. A full answer to Q4 further requires a description of the complete supervised ML procedure and the algorithms that have been used to optimize *S*. This description can be a pseudo-code of the algorithms or a mathematical description. In contrast, it is irrelevant for Q4 how exactly the algorithms are implemented, i.e., in which

programming language, with which libraries, etc., since this has no effect on their functionality.

Q5: Does *S* work like an alternative spam filter *S?**

This question aims at comparing two trained ML models that use the same set of input attributes to compute their outputs.

Suppose that *S* and *S** are white boxes, i.e., one knows their function types and parameters. Then *S* and *S** work equivalently if they use the same function type and have identical parameters. Note that the inverse is not true, i.e., two models can use different function types and parameters and still represent the same mathematical function. In this case, showing the equivalence of the two models is much more difficult.

However, if either *S* or *S** is a black box, one does not know any explicit representation of its function. In this case, one can only make equivalence statements with respect to a set of observations. For example, one can verify that *S* and *S** yield the same classification for a set of emails. This does not show, however, that the functions of *S* and *S** are identical—their results could differ on other observations.

Summing up, answering Q5 is simple if the answer is *no*: one only needs one email where *S* and *S** yield different classifications. In contrast, proving that the answer to Q5 is *yes* is difficult or almost impossible—except in the trivial case that both models are white boxes with the same function type and parameters.

Q6: Why does *S*'s classification differ from Alice's opinion?

This question aims at comparing Alice's personal reasoning with the classification function *S*. Answering the question requires investigating whether *S* relies on the same attributes of emails as Alice does. If this is not the case, one already has an answer: *S* relies on different attributes. If *S* and Alice rely on the same attributes, one must further compare the function *S* to the one of Alice, i.e., whether the attributes are used in the same way. We have already discussed this in the answer to Q5. Alice may also not be fully aware of her opinion, or might not be able to represent her reasoning sufficiently well to allow her to compare her personal reasoning to S. In this case, Q6 may not be answerable at all.[8]

Q7: What distinguishes spam emails from normal emails?

This question asks for a universal rule. If a commonly accepted definition of spam exists, then the answer is simple: if an email meets the definition of a spam email, then it is spam. Here the definition may be one reached by the method of conceptual analysis, but it might also be the case that it is just a stipulative definition that happens to be commonly accepted.

If the definition does not exist, one can take a set of examples of spam and normal emails and develop an ML model with it. The question then is whether the trained ML model is helpful in providing an answer to Q7. This depends on the degree of consensus among email recipients about what does and does not qualify as a spam email. If all recipients agree with each other, then any trained ML model with a high accuracy is a possible answer to Q7; and any such ML model could help to establish a commonly accepted definition of spam. However, if the recipients disagree on which email is spam or not, there can be no universal definition of spam, and Q7 is not answerable.

Questions addressed by XAI algorithms

In the following, we show that existing XAI algorithms do not give answers to any of the questions introduced in the previous section except for Q2. Each of the following paragraphs makes an observation about XAI algorithms that rules out some of the questions, until only Q2 remains. We then examine Q2 more closely and show the centrality of this question to present XAI research.

Real-world phenomena. There is consensus in the literature that XAI algorithms primarily explain ML models and not real-world phenomena (Gunning, 2017; Samek et al., 2017; Adadi and Berrada, 2018; Arrieta et al., 2020). Q7 asks for a characterization of a real-world phenomenon: spam. Although one can use ML models to explore possible answers to questions like Q7 (Sullivan, 2022), the decision of what the correct answer is, is independent of the models. Hence, Q7 is not addressed by XAI algorithms.

Creation process of ML models. XAI algorithms so far tend to generate "explanations" of ML models that are independent of how the models were created. To illustrate, consider a popular family of algorithms, so-called *feature attribution* methods, e.g., (Ribeiro et al., 2016; Sundararajan et al., 2017). Feature attribution methods estimate, for a given input to an ML model, how important the individual features of the input are for the corresponding output of the model. One can compute such feature attributions without knowing how the model was created. In contrast, knowledge of the creation process of the ML model *S* is necessary to answer why *S* classifies *M* as *spam* (Q1), and how *S* works (Q4).

We conclude that Q1 and Q4 are, at present, not answered by XAI algorithms.

Complex ML models. XAI research mainly focuses on ML models that are very complex, e.g., because they have millions of numeric parameters. In this setting, it is highly improbable that two ML models created in different ways end up with the same parameters and hence implement the same function. Thus, for the ML models considered by XAI research, Q5 and Q6 have trivial answers: Complex ML models hardly ever work in the same way; and the classification function of a user is expected to differ from that of a given complex ML model.

Q3 reduces to Q2. Recall that we have identified two ways to answer what distinguishes *spam* from *no spam* (Q3). The first applies traditional supervised ML methodology, namely feature engineering and the training of simple types of ML models.

This methodology was established long before XAI emerged as a research area. The second way of answering Q3 is to develop one or more complex ML models, like *S*, and then derive from them diverse simple functions that use interpreted attributes. This is equivalent to collecting different answers to Q2. Thus, if XAI algorithms address Q3, this is only partial, and only as a side effect of addressing Q2.

After these observations, only Q2 is left for discussion: how does *S* distinguish *spam* from *no spam*? For generality, we now decouple Q2 from the example of the thought experiment with ML model *S*. With the aim of discussing any kind of ML model using the methodology of supervised ML, we formulate a question that is applicable more broadly and deems it a "generalized" question:

Q*: How can one represent an ML model as a simple function that uses interpreted attributes?

From our perspective, this is the core question addressed by XAI algorithms. Almost all existing XAI algorithms contribute in some way to approximating complex ML models with simpler functions (e.g., Guidotti et al., 2019), or with functions that use interpreted instead of technical attributes (e.g., Bau et al., 2017; Kim et al., 2018; Linardatos et al., 2021). Considering the ambiguity of terms that the literature usually relies on to describe the capabilities of XAI algorithms—i.e., *interpretability*, *explainability*, and so on—it seems surprising that one can accurately describe these capabilities with one simple question. We have the impression that literature and debates often make normative

claims on XAI algorithms that go far beyond the question Q*. For example, it is far from clear how Q* relates to terms like "trust", or "fairness", which carry normative connotations. While these normative terms may relate to XAI algorithms in one or another way, using them without an explicit and shared interpretation obfuscates the results of XAI research. It spurs misconceptions on the capabilities of XAI algorithms in politics and society, and ultimately harms the credibility of ML research. So, we advocate a more realistic approach to describe the capabilities of XAI algorithms: by focusing on the questions they can currently address.

Challenges for XAI algorithms

Given the core question Q* addressed by XAI algorithms, we now examine what is difficult about answering it. We identify two main challenges.

Approximation. The first challenge is to approximate complex ML models with simpler functions, so-called *surrogate models*.[9]

An elementary way to obtain a surrogate model is to observe a large set of inputs and corresponding outputs of the complex ML model and then train a simple ML model on these observations. That is, the surrogate model is optimized to predict the outputs of the complex model, and not to predict the "true" outputs. The accuracy of the surrogate model with respect to the complex model, i.e., how well the surrogate predicts the complex model, is sometimes called *fidelity* (Guidotti et al., 2019).

Approximating ML models is both a technical and a conceptual challenge. For instance, there is a tradeoff between the fidelity and the complexity of surrogate models: Simpler surrogate models will, in general, achieve lower fidelity than complex ones. Thus, which surrogate model is optimal for users depends on how the users balance fidelity and complexity.

Further, there are many definitions of the complexity of an ML model. One way is to count the number of computational steps that an algorithm performs to obtain the output of the model for a given input. However, there are also definitions that are specific to types of ML models. For instance, one often quantifies the complexity of decision trees, a specific type of ML model, with metrics that are specific to tree structures, e.g., the length of the longest path from the root to a leaf node. The lack of a universal definition of model complexity can make it difficult to choose between different types of surrogate models.

Translation. The second challenge in answering Q* is to translate the technical attributes that ML models use to discriminate between data items into *interpreted attributes*. Technical attributes cover both the inputs of an ML model and the intermediate results that the model computes for its prediction. Interpreted attributes are all attributes of data items that humans can interpret and communicate to each other using common terminology in a shared language, see the elaborations on Q2 in the section "Questions about ML models".

The difficulty of the translation challenge depends on the design and application of the ML model. For some models and applications, it is clear how interpreted attributes relate to the technical attributes used by the model. For example, suppose that the spam filter S uses a digital representation of emails that consists of a single bit, i.e., a 0 or 1.

Suppose further that this bit is set if and only if an email contains the word "money".

In this case, the technical attribute has a trivial mapping to an interpreted attribute. In other cases, this mapping exists but is unknown. For example, suppose that S relies only on the first principal component of an email vector embedding and that the value of this principal component correlates strongly with the presence of the word "money". One does not know this correlation until investigating it. Finally, for some models and applications, no terminology exists to accurately describe the inputs or the intermediate results of the models. In the previous example, this is if the principal component used by the ML model does not correlate with any known term used to describe emails. In summary, the translation challenge can be trivial, difficult, or even unsolvable, depending on the ML application.

The state of XAI algorithms

We now assess how far existing XAI algorithms address the approximation and translation challenges raised by question Q*. This section is not a comprehensive review, but rather a summary of present XAI algorithms with some examples.

There is a plethora of XAI algorithms that address the approximation challenge, for different types of complex ML models and surrogate models, and for different definitions of complexity and fidelity. To give some examples, some algorithms can approximate complex ML models, such as neural networks and random forests, with simpler models, such as decision trees (Craven and Shavlik, 1995; Bastani et al., 2019) or rule lists (Bénard et al., 2021). There are also algorithms to approximate complex ML models locally, i.e., in the vicinity of a given input. Popular examples of this are LIME (Ribeiro et al., 2016), SHAP (Lundberg and Lee, 2017) and integrated gradients (Sundararajan et al., 2017). See Guidotti et al. (2019) for a comprehensive overview of approximation methods.

There is also a philosophical account that uses the language of approximation (Erasmus et al., 2021). Prima facie, this account by Erasmus et al. might be seen as close to what we propose here. Especially, as their account is focused on interpretation, one might think that our notion of translation is another formulation for their envisioned interpretation. However, Erasmus et al. conceptualize interpretation as a relation between an "interpretans" and an "interpretandum", and state that "both the interpretans and the interpretandum are explanations" (2021, 851). Additionally, they also claim that approximation is related to interpretation and by that is related to explanations understood as interpretans and interpretandum (cf. 2021, 853f.). We see approximation differently, namely as the development of a surrogate model that is not necessarily interpreted. In the following, we hence use the term approximation but do not speak of interpretation in the context of approximation.

Compared to the approximation challenge, the translation challenge has received less attention. Some work addresses ML models for computer vision, i.e., convolutional neural networks (CNNs) that make predictions based on images. There are XAI algorithms to discover correlations of neurons or layers in CNNs with interpreted attributes of images, say, visible objects, shapes, or colors (Zhou et al., 2019; Kim et al., 2018; Cammarata et al., 2020; Goh et al., 2021). The discovered correlations enable users to map technical attributes used by CNNs, i.e., neuron or layer activations, to interpreted attributes. The main limitation of these methods is that they require a comprehensive database of annotated concepts on images.

To address this issue, Ghorbani et al. (2019) extract concepts in an automated way. Experiments in this work indicate that many technical attributes of images used by neural networks can be mapped to interpreted ones. Similar to these examples from computer vision, there is also work on translating ML models for natural language processing (Poerner et al., 2018) and speech recognition (Krug et al., 2018).

Answering question Q* often poses a combination of approximation and translation challenges. Think of a typical

computer vision model, an image classifier. The classifier uses technical attributes of images, typically numeric pixel values for each color channel of the image. Now suppose that one approximates the classifier with a surrogate model. The surrogate model will either have low fidelity, or it will be almost as complex as the classifier. The reason is that image classification is an inherently complex task:

Functions that map from pixel-wise color values to the target classes are in most cases either complex or not accurate. However, an answer to Q* can ignore some of this inherent complexity. This is because human users think and speak about images with interpreted attributes, e.g., shapes or objects. Answering question Q* does not require approximation of any parts of an image classifier that serve only to reconstruct interpreted attributes from pixel values. On the contrary: a better answer to Q* is one that (i) translates technical attributes in the classifier to interpreted ones, and (ii) defines an approximation of the classifier based on the translated attributes.

Such joint translation and approximation of ML models is an open challenge for XAI algorithms. It occurs in all applications of ML models where the inputs have a lower level of abstraction than the terminology that users rely on. Besides computer vision, this also holds for speech recognition and natural language processing. Another example is the application of ML models to predict physical processes; e.g., ML models that derive their predictions from the states of individual atoms or molecules, whereas their users observe aggregated changes on the macro level of the material under study. One way to address the joint approximation and translation challenge is *neurosymbolic AI*, i.e., "to develop neural network models with a symbolic interpretation" (Garcez and Lamb, 2020). A pioneering example is the Neural Prototype Tree (Nauta et al., 2021), i.e., a neural network that learns a few "prototypical" combinations of interpreted attributes and then classifies inputs based on their resemblance to these prototypes.

Conclusions

A problematic reasoning scheme in the literature currently obfuscates the relation between general goals and capabilities of existing XAI algorithms. In this paper, we have explored another way to characterize XAI algorithms, namely from the perspective of their users. We have found that current XAI algorithms primarily address one particular question that users have in the context of ML, by disambiguating, re-interpreting, and generalizing this question to: "How can one represent a complex ML model as a simple function of interpreted attributes?".

The succinctness of this question contrasts with the ambiguity of terms used in the literature to characterize XAI algorithms, such as *interpretability* and *explainability*. Other terms in the literature have a normative connotation, e.g., *trust*, whereas the identified question is purely technical. The contrast between prevalent terminology and the actual goals of XAI algorithms may spur excessive expectations of the algorithms on the part of policymakers and society. To avoid this, we propose to focus on questions that the algorithms can actually help with. Our analysis of XAI algorithms further reveals two key challenges for XAI research: the approximation and translation of ML models. Regarding approximation, the literature already offers many approaches; slightly fewer approaches exist for translation. We think that holistic methods that address both challenges will be key in future research.

Data availability

The research didn't involve any empirical studies as the paper is mainly a philosophical contribution. So, there aren't any empirical data to be shared. The authors are happy to share other kinds of data based on personal requests.

Received: 26 April 2023; Accepted: 6 June 2024;
Published online: 14 June 2024

Notes

1 The term "explanation" itself has multiple interpretations and has been a source of controversial debates. We elaborate on different perspectives from philosophy and computer science as regards explanation in the section "Ambiguities of the questions". For a criticism of the focus on explanation and an alternative approach that speaks in terms of understanding rather than in terms of explanation see Páez (2019). We follow the common way of speaking about XAI algorithms according to which they *provide* explanations. However, we do not wish to enter the debate about which philosophical account of explanation is preferable; the interested reader may consult, e.g., Woodward and Ross (2021) for explanation, in general, or Erasmus et al. (2021) that discusses different accounts of explanation in relation to the task of explaining artificial neural networks. In particular, one may ask whether the accounts proposed for "scientific explanation" also fit XAI, or whether a specific account of explanation is needed. We further build on the intuition that the aim of a user is to understand ML models. Whether this understanding is to be cashed in with an account of explanatory understanding or an account of objectual understanding is another issue that deserves its own debate, see Hills (2016); Baumberger et al. (2017); Baumberger and Brun (2020) for details. Our approach is in line with the proposal of Fleisher (2022) to take understanding as a more basic notion than an explanation. However, Fleisher (2022) argues for a philosophical account of understanding. Here, we instead argue from the practice of computer science and suggest a new analysis of challenges for XAI research based on a philosophical and interdisciplinary perspective.

2 We define these two modest capabilities in the context of our discussion of question Q2 in the section "Questions about ML models".

3 To give the reader a first glimpse into what could be meant by crucial terms such as 'interpretable' or 'explainable', we cite definitions by Will Fleisher and Adrian Erasmus and collaborators. The first one reads: "A model is interpretable for a stakeholder iff it is understandable by that stakeholder (to a sufficient degree)." And, "A model is explainable to a stakeholder iff it is accompanied by an XAI method whose output provides an explanation that puts the stakeholder in a position to understand why the system made the decision it did" (Fleisher, 2022, p. 12). An alternative definition is given by Erasmus et al. that is based on 'interpretation' and 'explanation': "interpretation is something that one does to an explanation to make it more understandable," and "interpretability [is] […] the ability to provide an interpretation" (Erasmus et al., 2021, p. 849).

4 In our discussion, we left it open whether trust is more or less equivalent to a form of reliance or not. One may also ask if trust must be systematically related to trustworthiness (cf. Nickel, 2021).

5 Beside deductive nomological explanations also causal or mechanist explanations might be further options to disambiguate the question. For example, in Erasmus et al. (2021) these three accounts of explanation and an additional fourth one (the account of inductive statistical explanation) are discussed in detail.

6 We thank an anonymous reviewer for making us aware of the special nature of this question.

7 Khalifa defines generic understanding in the following way: "S has some understanding of why p if and only if 'S understands why p' is true in some context C" (Khalifa, 2017, p. 5).

8 Thanks to an anonymous reviewer for pointing this out.

9 A technical remark: Our broad definition of "surrogate model" includes approaches such as feature attribution or counterfactuals, because one can view them as local approximations of an ML model. The usage of this term may differ in other literature on XAI.

References

Abdi H, Williams LJ (2010) Principal component analysis. WIREs Comput Stat 2(4):433–459. https://doi.org/10.1002/wics.101

Adadi A, Berrada M (2018) Peeking inside the Black-Box: a survey on Explainable Artificial Intelligence (XAI). IEEE Access 6:52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

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

Bastani O, Kim C, Bastani H (2019) Interpreting Blackbox models via model extraction. arXiv. https://doi.org/10.48550/arXiv.1705.08504

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

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

Baumberger C, Brun G (2020) Reflective equilibrium and understanding. Synthese 198:7923–7947. https://doi.org/10.1007/s11229-020-02556-9

Bénard C, Biau G, da Veiga S, Scornet E (2021) Interpretable random forests via rule extraction. In: Banerjee A, Fukumizu K (eds) Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, vol 130. PMLR, pp. 937–945. https://proceedings.mlr.press/v130/benard21a.html

Cammarata N, Goh G, Carter S, Schubert L, Petrov M, Olah C (2020) Curve detectors. Distill 5(6):e00024.003. https://doi.org/10.23915/distill.00024.003

Craven M, Shavlik J (1995) Extracting tree-structured representations of trained networks. In: Touretzky D, Mozer MC, Hasselmo M (eds) Advances in neural information processing systems, vol 8. MIT Press. https://proceedings.neurips.cc/paper/1995/file/45f31d16b1058d586fc3be7207b58053-Paper.pdf

Erasmus A, Brunet TDP, Fisher E (2021) What is interpretability? Philos Technol 34(4):833–862. https://doi.org/10.1007/s13347-020-00435-2

Fleisher W (2022) Understanding, idealization, and explainable AI. Episteme:1–27. https://doi.org/10.1017/epi.2022.39

Garcez A. d'Avila, Lamb LC (2020) Neurosymbolic AI: The 3rd Wave. arXiv. https://doi.org/10.48550/arXiv.2012.05876

Ghorbani A, Wexler J, Zou J, Kim B (2019) Towards automatic concept-based explanations. arXiv. https://doi.org/10.48550/arXiv.1902.03129

Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L (2018) Explaining explanations: an overview of interpretability of machine learning. Paper presented at the IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), IEEE. pp. 80–89

Goh G, Cammarata C, Voss C, Carter S, Petrov M, Schubert L, Radford A, Olah C (2021) Multimodal neurons in artificial neural networks. Distill 6(3):e30. https://doi.org/10.23915/distill.00030

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

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2019) A survey of methods for explaining Black Box models. ACM Comput Surv 51(5). https://doi.org/10.1145/3236009

Gunning D (2017) Explainable Artificial Intelligence, Defense Advanced Research Project Agency. https://www.darpa.mil/program/explainable-artificial-intelligence. Accessed 13 Jun 2022

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

Hoffman RR, Mueller ST, Klein G, Litman J (2018) Metrics for explainable AI: challenges and prospects. arXiv https://arxiv.org/abs/1812.04608

Khalifa, K (2017) Understanding, explanation, and scientific knowledge. Cambridge University Press, New York. https://doi.org/10.1017/9781108164276

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: Proceedings of the 35th International Conference on Machine Learning. Presented at the International Conference on Machine Learning. PMLR, pp. 2668–2677. https://proceedings.mlr.press/v80/kim18d.html

Krishnan M (2020) Against interpretability: a critical examination of the interpretability problem in machine learning. Philos Technol 33(3):487–502. https://doi.org/10.1007/s13347-019-00372-9

Krug A, Knaebel R, Stober S (2018) Neuron activation profiles for interpreting convolutional speech recognition models. Paper presented at the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada, pp. 1–13. https://openreview.net/pdf?id=Bylpgfjen7

Le Q, Mikolov T (2014) Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning. Presented at the International Conference on Machine Learning. PMLR, pp. 1188–1196. https://proceedings.mlr.press/v32/le14.html

Liao Q, Gruen D, Miller S (2020) Questioning the AI: informing design practices for explainable AI user experiences. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp. 1–15. https://doi.org/10.1145/3313831.3376590

Linardatos P, Papastefanopoulos V, Kotsiantis S (2021) Explainable AI: a review of machine learning interpretability methods. Entropy 23(1). https://doi.org/10.3390/e23010018

Lipton ZC (2018) The mythos of model interpretability. Commun ACM 61(10):36–43. https://doi.org/10.1145/3233231

Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, pp. 1–10. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Mittelstadt B, Russell C, Wachter S (2019) Proceedings of the Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA. FAT* '19, p. 279288. https://doi.org/10.1145/3287560.3287574

Molnar C (2020) Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/

Nauta M, van Bree R, Seifert C (2021) Neural prototype trees for interpretable fine-grained image recognition. arXiv. https://doi.org/10.48550/arXiv.2012.02046

Nickel PJ (2021) Trust in engineering. In: Michelfelder D, Doorn N (eds) Routledge handbook of the philosophy of engineering. Routledge, New York, pp. 494–505

Páez A (2019) The pragmatic turn in Explainable Artificial Intelligence (XAI). Mind Mach 29(3):441–459

Poerner N, Roth B, Schütze H (2018) Interpretable textual neuron representations for NLP. In: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: analyzing and interpreting neural networks for NLP. Association for Computational Linguistics, Brussels, Belgium, pp. 325–327

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. Association for Computing Machinery, New York, NY, USA, pp. 1135–1144

Robbins S (2019) A misdirected principle with a catch: explicability for AI. Minds Mach 29(4):495–514. https://doi.org/10.1007/s11023-019-09509-3

Samek W, Wiegand T, Müller K-R (2017) Explainable Artificial Intelligence: understanding, visualizing and interpreting deep learning models. arXiv. https://doi.org/10.48550/arXiv.1708.08296

Sullivan E (2022) Understanding from machine learning models. Br J Philos Sci 73(1):109–133. https://doi.org/10.1093/bjps/axz035

Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia. PMLR, 70, pp. 1–10

Szczepański M, Choraś M, Pawlicki M, Pawlicka A (2021) The methods and approaches of Explainable Artificial Intelligence. In: Paszynski M, Kranzlmüller D, Krzhizhanovskaya VV, Dongarra JJ, Sloot PMA (eds) Computational science—ICCS 2021. Springer International Publishing, Cham, pp. 3–17

Tomsett R, Braines D, Harborne D, Preece A, Chakraborty S (2018) Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. arXiv. https://doi.org/10.48550/arXiv.1806.07552

Verma S, Boonsanong V, Hoang M, Hines KE, Dickerson JP, Shah C (2022) Counterfactual explanations and algorithmic recourses for machine learning: a review. arXiv. https://doi.org/10.48550/arXiv.2010.10596

Woodward J, Ross L (2021) Scientific explanation. In: Zalta EN (ed.) The Stanford Encyclopedia of Philosophy (Summer 2021 Edition). Metaphysics Research Lab, Stanford University https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/

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 B, Bau D, Oliva A, Torralba A (2019) Comparing the interpretability of deep networks via network dissection. In: Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R (eds) Explainable AI: interpreting, explaining and visualizing deep learning. Springer International Publishing, Cham, pp. 243–252

Author contributions

All authors—MR, HT, MP, and RH—contributed to the conception and design of the article. The first development of the thought experiment was done by HT. An initial draft of the manuscript was written by MR, and all authors commented on and contributed to previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

Informed consent was not required as the study did not involve human participants.

Additional information

Correspondence and requests for materials should be addressed to Michael Poznic.

Reprints and permission information is available at http://www.nature.com/reprints

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