A Means-End Account of Explainable Artificial Intelligence

Oliver Buchholz

Received: 27 January 2023 / Accepted: 29 June 2023 / Published online: 17 July 2023
© The Author(s) 2023

Abstract

Explainable artificial intelligence (XAI) seeks to produce explanations for those machine learning methods which are deemed opaque. However, there is considerable disagreement about what this means and how to achieve it. Authors disagree on what should be explained (topic), to whom something should be explained (stakeholder), how something should be explained (instrument), and why something should be explained (goal). In this paper, I employ insights from means-end epistemology to structure the field. According to means-end epistemology, different means ought to be rationally adopted to achieve different epistemic ends. Applied to XAI, different topics, stakeholders, and goals thus require different instruments. I call this the means-end account of XAI. The means-end account has a descriptive and a normative component: on the one hand, I show how the specific means-end relations give rise to a taxonomy of existing contributions to the field of XAI; on the other hand, I argue that the suitability of XAI methods can be assessed by analyzing whether they are prescribed by a given topic, stakeholder, and goal.

Keywords XAI · Explainability · Means-end epistemology · Epistemic normativity · Instrumental rationality

1 Introduction

Methods of machine learning (ML) are gaining relevance in a variety of domains. They are employed to navigate self-driving vehicles (Bojarski et al., 2016), support medical diagnosis (Esteva et al., 2017), and to detect objects ranging from particles at...
the subatomic level (Baldi et al., 2014) to exoplanets (Foreman-Mackey et al., 2015).¹

At the same time, it is widely acknowledged that their very nature makes ML methods essentially ‘black boxes’ to human agents. Both their specific functioning and their complexity, especially in the field of deep learning, are inherently opaque or beyond human grasp (Burrell, 2016; Creel, 2020).² Thus, frequently, laymen and experts alike do not know how these methods work, why they are successful, or why they fail. This observation led to the widely shared belief that ML methods should be made explainable and sparked research in the field of explainable artificial intelligence (XAI). Since the ‘X’ in ‘XAI’ stands for ‘explainable’, it seems straightforward and uncontroversial to summarize the central aim of the field plain and simple as follows: XAI seeks to provide instruments that produce explanations of ML methods.³

Yet on the one hand, there is profound conceptual disagreement regarding the precise meaning of this aim (Lipton, 2018): what is it that explanations of ML methods should achieve? Suggestions in the literature comprise the straightforward requirements of explainability and interpretability (Doshi-Velez & Kim, 2017; Erasmus et al., 2021), the more intuitive understandability (Páez, 2019) as well as the larger category of transparency (Günther & Kasirzadeh, 2021; Zerilli et al., 2019). Others argue that all these notions lack precise definitions altogether and only hide the real goals such as safety of or non-discrimination by ML methods (Krishnan, 2020).

On the other hand, the literature also diverges with respect to the instruments that are deemed appropriate to achieve the aim of XAI. Popular approaches range from local approximations of complex models (Ribeiro et al., 2016) to visual (Xu et al., 2015), textual (Hendricks et al., 2016), counterfactual-based (Wachter et al., 2018) as well as attribution-based explanations (Lundberg & Lee, 2017). Overall, it is unclear whether the field of XAI is scattered into various disconnected subprojects or whether there is a common structure that is shared by the variety of approaches pursued in the literature.

In this article, I propose an account that structures the field. To do so, I employ insights from means-end epistemology.⁴ Means-end epistemology takes epistemology to be a normative discipline; it is based on the principle of instrumental rationality (Huber, 2021, p. 1). Rational agents are assumed to have certain epistemic ends (e.g., the end to hold true beliefs), and such agents ought to adopt certain means (e.g., to base their beliefs on total evidence) if and only if they further these epistemic ends (Huber,

---

¹ I use the term ‘ML method’ to refer to some general methodology, e.g., deep learning, and distinguish it from the term ‘ML model’ by which I refer to a specific model, e.g., a deep neural network for which the weights have already been determined. Thus, the term ‘ML method’ explicitly includes the learning process, while the term ‘ML model’ only refers to the learned function.

² The qualification ‘inherently’ is meant to emphasize the distinction highlighted by Schubbach (2021) between an in-fact-impossibility to grasp the functioning of technical objects due to a lack of time or expertise and an in-principle-impossibility to grasp the functioning of ML methods due to their functioning being established through a learning process.

³ To be precise, this applies to the field insofar it is concerned with the application of XAI techniques to opaque ML methods, which is the focus of this paper. It does not directly apply to that part of XAI that seeks to develop and apply interpretable methods, thereby avoiding the use of opaque ML methods from the outset (Rudin, 2019).

⁴ The idea of exploiting existing research from philosophy and the social sciences when analyzing XAI is put forward, e.g., in Miller (2019).
2021, p. 125; Schulte, 1999, p. 7). Providing an explanation is such an epistemic end, because it relates to the epistemic state of the agent receiving the explanation.\(^5\) Thus, when trying to provide instruments that produce explanations of ML methods, the field of XAI essentially seeks to provide the means to further an epistemic end. Consequently, means-end epistemology is a very suitable framework to analyze XAI. I call this the means-end account of XAI. The account’s motivation is based on the fundamental observation that means-end relations are crucial to the field of XAI. Indeed, as I will point out, researchers explicitly or implicitly specify them in their contributions. I will show how these means-end relations can be specified on a fine-grained level by distinguishing what should be explained (topic), to whom something should be explained (stakeholder), why something should be explained (goal), and how something should be explained (instrument).

I will argue that overall, the means-end account has several important consequences. First, it explains why disagreement arises in the field: the divergence in methods of XAI follows from the disagreement on fine-grained ends being pursued. Second, it unifies and structures the field: there is a shared methodology of developing appropriate means for given ends. Third, this structure has a descriptive component: different authors specify different topics, address different stakeholders, and have different goals. Thus, they pursue different ends and different means are appropriate to achieve them. This gives rise to a taxonomy that classifies existing contributions to the field along the specific means-end relations that are considered. Fourth, the means-end structure also has a normative component: according to means-end epistemology, different means ought to be rationally adopted to achieve different epistemic ends. Therefore, the fine-grained ends of an explanation normatively constrain the set of admissible means to achieve it. The means-end account thus reveals how the suitability of particular instruments of XAI is prescribed by the ends for which an explanation is sought. I argue that this analysis gives rise to a normative framework that can be used to assess the suitability of XAI techniques.

Consequently, this article extends that strand of the literature that emphasizes the relevance of pragmatic considerations in the context of XAI.\(^6\) Indeed, I show that pragmatic considerations are relevant, especially in what concerns the topics, stakeholders, and goals. However, by focusing on the principle of instrumental rationality as a cornerstone of epistemic normativity, I additionally provide a normative justification for why pragmatic considerations ought to be relevant. Furthermore, while I do not claim to be the first developing a normative framework for XAI, this article is the first to thoroughly base such a framework on an underlying epistemological theory.\(^7\)

The remainder of this article is structured as follows: In Sect. 2, I outline the main characteristics of XAI by providing concrete examples from the field. I also give an introduction to means-end epistemology. In Sect. 3, I argue that problems of XAI can be framed as problems of means-end epistemology. I thereby establish the means-

---

5 This is also a well-established result in psychology as reported, for instance, by Keil (2006) or Lombrozo (2006).
6 Examples include Besold and Uckelmann (2018), Nyrup and Robinson (2022), and Páez (2019).
7 Other normative frameworks for XAI have been proposed, for instance, by Langer et al. (2021), Mohseni et al. (2021), and Zednik (2021).
end account of XAI and discuss both its normative and its descriptive component. In Sect. 4, I discuss an extension of the account before concluding in Sect. 5.

2 Preliminaries

This section outlines the main characteristics of XAI by examining two examples from the field. It also gives a short introduction to those aspects of means-end epistemology that are relevant for the rest of the paper.

2.1 Explainable artificial intelligence

The field of XAI is concerned with developing techniques that are commonly employed to eliminate the opacity of many ML methods. In a first step, there is usually some ML model, for instance a complex deep neural network, that is deemed opaque. In a second step, some XAI technique is applied to that method to (partly) eliminate the perceived opacity. Yet as straightforward as it may seem, there is considerable disagreement about what ‘eliminating opacity’ amounts to. This disagreement is perhaps best illustrated by outlining two of the most well-known methods of XAI.

First, consider local interpretable model-agnostic explanations (LIME, Ribeiro et al., 2016). Explanations produced by LIME are local in the sense that they explain ML predictions only within a specific part of the data. For instance, a complicated ML model might be approximated locally by a much more intuitive linear function. This makes the model locally interpretable, because the approximation reveals its behavior within a specific part of the data in a human-graspable way. Since LIME is model-agnostic, it is applicable to a wide range of ML methods. What matters is that their learning process is completed, because LIME is applied to the final ML model.

Second, consider counterfactual explanations (CFE, Wachter et al., 2018). Just as LIME, they are targeted at explaining the final ML model that becomes available after the learning process. More precisely, CFE focus on the input data of the model and evaluate which changes in which input feature would lead to a change in the predicted output. The outcome of this evaluation is usually provided as a verbalized counterfactual statement that takes the following form: “You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan.” Thus, CFE provide reasons for a given decision and seem to specify how to reach a desired result in the future. As becomes obvious, CFE exclusively focus on a model’s input-output relation: they explain, for a single case, why a certain input led to a certain output and they do so “without opening the black box”, that is, without having a closer look at the characteristics of the given model. They explain the model’s output by giving seemingly intuitive reasons rather than by detailing the mathematical relationships that actually led to the output at hand.

Yet research has shown that there are nevertheless situations in which LIME can fail (Garreau & von Luxburg, 2020).

I use the rather cautious ‘seem’, since this view is controversial. See Sect. 3.3 for a brief overview about the debate in XAI. See also Raidl and Rott (forthcoming) as well as references therein for the general philosophical debate about the relation of counterfactuals and ‘because’-statements.
Overall, the deliberately simplistic sketch of two of the most popular methods in XAI illustrates the observation from above: that instruments of XAI differ widely and that there is no consensus on which instruments are appropriate to use. This raises the question whether there is nevertheless something that XAI techniques, ranging from LIME to CFE, have in common. As a first step, let us try to distill an overall aim of XAI, general enough to be shared by LIME, CFE, and other XAI techniques as different as they may be: what is it, that XAI tries to achieve?

First, XAI develops instruments, for instance, LIME or CFE. Second, these instruments are meant to produce explanations of the ML methods to which they are applied. Thus at the highest level, the aim of XAI might be summarized by stating that XAI seeks to provide instruments that produce explanations of ML methods.

Clearly, the term ‘explanations’ in (AIM) is dangerously loaded, both within philosophy and—partly following from that—within XAI. Mainly in the philosophy of science, there is an extensive debate about the criteria for a proper explanation. This debate is increasingly being taken up in the XAI literature to assess whether and to what extent said criteria also apply to explanations of ML methods. There is no doubt that I will be unable to settle the philosophical debate on explanation in the present paper. I will thus not take a specific stand on it either and instead rely on a single, sufficiently uncontroversial insight: there is a particular way of looking at the issue of explanation and at different accounts of explanation proposed in the literature that Salmon (1984, Ch. 4) calls the epistemic conception of explanation. As its name suggests, the focus of this conception is on the epistemology of explanation, that is, on the relation between explanations and an individual who is asking for, constructing, or receiving them. Seen that way, the internal structure of explanations that is a focal point in the overall debate mainly matters with regards to the needs of the respective stakeholder, not for merely logical reasons alone.

With this epistemic perspective in place, my analysis of XAI will proceed as follows: at this stage, my sole aim is to ensure a maximally general formulation of the high-level aim of XAI that is meant to be as uncontroversial as possible. Thus, in (AIM), I use the term ‘explanations’ to designate the different kinds of outputs that are produced by XAI techniques, regardless of whether they qualify as a proper explanation according to some account or not. Below, when putting forward the main argument, I will return to the issue and argue that what counts as a proper explanation in the context of XAI should mainly depend on epistemic considerations (see p. 19).

---

10 The classical overview of this debate is Salmon (1989). For a concise yet comprehensive summary, see Woodward and Ross (2021).

11 For instance, Erasmus et al. (2021) discuss the deductive-nomological as well as the inductive-statistical account of explanation in the context of XAI, while Räz (2022) focuses on the statistical-relevance account.

12 For two exemplary contributions to this literature, see Gärdenfors (1980) and Potochnik (2016).

13 To be precise, formulations such as ‘XAI develops’ or ‘the aim of XAI’ are shorthands that equate XAI as a field with the people working on or applying XAI.
2.2 Means-end epistemology

The second building block of the account that I propose in this paper is means-end epistemology. It takes epistemology to be a normative discipline. In this context, normativity has an epistemic rather than a moral meaning. Contrary to moral normativity that is concerned with what one ought to do, epistemic normativity is concerned with what one ought to believe. In the case of means-end epistemology, the central normative criterion is spelt out by the principle of instrumental rationality (Huber, 2021; Schulte, 1999):

Given an epistemic end, means ought to be adopted if and only if they further the epistemic end.

(NORM)

Thus, under the assumption that agents can have certain epistemic ends, the principle of instrumental rationality requires them to adopt those means that are appropriate to achieve the given end(s). For instance, consider an agent who pursues the epistemic end of holding only true beliefs. They falsely believe that they were denied a loan because their annual salary is too low. Now suppose that they receive the (true) information that, in fact, they were denied the loan because their loan application contained transposed digits for the annual salary. According to means-end epistemology, the agent—if rational—ought to adopt means that lead to a revision of their initial and false belief in favor of the true belief that their loan application contained transposed digits for the annual salary, since these means would further their epistemic end of holding only true beliefs.

The strategy of equating epistemic normativity with instrumental rationality is a widely held position in epistemology and some well-known approaches follow this route without explicitly carrying the label ‘means-end epistemology’. A prominent example are accuracy arguments for probabilism according to which a rational agent ought to structure their beliefs in line with the axioms of probability given the end of maximally accurate beliefs (Joyce, 1998). Yet equating epistemic normativity with instrumental rationality and expressing it in terms of means-end relationships as in (NORM) also involves several aspects that are important for the subsequent discussion of XAI.

First, different ends call for different means. For instance, Schulte (1999) conducts a means-end analysis of inductive inference to determine which inductive method an agent ought to adopt given a particular epistemic end. He shows that the end of the method and, hence, the agent’s beliefs converging to the truth with minimal retractions conflicts with the end of time-minimal convergence to the truth. Most importantly,
however, he also shows that each of the ends calls for a different inductive method that is ideal for achieving it (Schulte 1999, p. 20). Thus, generally speaking, while the adoption of certain means might be rational for agents with certain epistemic ends, agents with different ends need not or even ought not adopt the same means (Schulte 1999, p. 26). This is a consequence of the normative perspective adopted by means-end epistemology.

Second, note that ‘certain means’ in (NORM) generally refers to a set of means, $M$, rather than to one particular means, $m \in M$. Thus, different ends may call for different sets of means, $M$ and $M'$, but the latter can be partially overlapping, such that $M \cap M' \neq \emptyset$. So on the one hand, different ends might be achieved by the same particular means $m$. On the other hand, the same end might be achieved by different particular means, $m$ and $m'$.

Third, epistemic ends come in different levels of granularity. For instance, the starting point of Schulte’s (1999) means-end analysis of inductive inference is the epistemic end of beliefs converging to the truth. Subsequently, however, further details are added to this initial end by requiring the convergence to possess certain properties, most importantly to evolve fast or with minimal retractions. Clearly, the latter ends are more fine-grained than the former, thereby allowing to identify a more specific set of means that ought to be rationally adopted. This insight also suggests that it is preferable to analyze epistemic ends separately rather than as a set if there are several of them: analyzing a single (even coarse-grained) end is more specific than analyzing a set of distinct ends which is why it makes the hypothetical part of (NORM), ‘given an epistemic end’, more precise. As a consequence, it also allows one to derive a more precise imperative stating the means that ought to be adopted.

Fourth, note that the means-end statement in (NORM) has the logical form of a biconditional and is therefore equivalent to a conjunction of two conditional statements pointing in two opposite ‘directions’: according to one direction, if certain means ought to be adopted, then they further the epistemic end; according to the other and, arguably, more natural direction, if they further the epistemic end, then certain means ought to be adopted. The importance of taking both directions into account and thus of preserving the logical form of the biconditional will become important below.

3 A means-end account of XAI

This section establishes what I shall call the means-end account of XAI. It also discusses the account’s descriptive and normative component.

3.1 Establishing the account

Recall the overall goal of XAI that I proposed in (AIM): XAI seeks to provide instruments that produce explanations of ML methods. On the one hand, this means that in virtue of producing explanations, XAI instruments are clearly targeting the epistemic states of agents, be it their beliefs about or their understanding of ML methods—the observation does not hinge on a particular epistemological concept such as belief,
knowledge, or understanding being affected. Consequently, it is fair to say that the
development and use of XAI are closely tied to the pursuit of epistemic ends. On the
other hand, XAI seeks to provide the instruments to achieve such epistemic ends. Put
differently, the development and use of XAI are essentially about finding appropriate
means to further given epistemic ends. Taken together, these aspects reveal that

Problems of XAI can be framed as problems of
means-end epistemology. (FRAME)

One might object that it is somewhat circular to argue for (FRAME) based on the
high-level goal of XAI proposed in (AIM), since the latter could have been purposefully
designed for precisely this argument. However, as mentioned above, (AIM) is simply
formulated on the highest level possible so as to give rise to a maximally broad and
uncontroversial statement.

In a nutshell then, the insight in (FRAME) means that developers or users of XAI
can be conceived of as agents that pursue some specific epistemic end and that try to
come up with or apply the appropriate XAI instrument for achieving the given end.
Let us return to the examples from Sect. 2.1 for illustration. First, consider LIME: as
outlined above, LIME seeks to explain an individual prediction or several predictions
in a specific region of the data. This constitutes the epistemic end that is pursued in this
case. The means proposed to achieve this end are local approximations to a possibly
complex ML model. Second, consider CFE: here, the epistemic end is to explain the
input-output relation of a model. The means proposed in this case are counterfactual
statements like the one about loan denial given above. So indeed, means-end relations
seem useful and easily applicable to analyze XAI techniques. It is for this reason that
I am adopting the perspective of means-end epistemology in this article, not because
I consider epistemological debates about that perspective or the related one about
epistemic consequentialism to be settled.17

However, when framing problems of XAI as problems of means-end epistemology,
it might seem trivial to proceed by simply specifying means and ends corresponding
to existing methods like LIME or CFE. The more important question is how means-
end relations are determined in practice, for instance, in the development of new
methods. Analyzing the means-end relations governing XAI more closely reveals that
researchers either explicitly or implicitly answer a variety of different questions: what
should be explained, to whom it should be explained, and how it should be explained.
I argue that the answers that are specified for each of the questions determine the
relevant means-end relations.

Answering the question as to what should be explained determines the topic of the
explanation. At first blush, this might seem unnecessary: when trying to produce an
explanation of some ML method, the topic should obviously be the ML method itself.
Yet the topic can also be spelled out at a more fine-grained level as a particular aspect
of an ML method and, indeed, it regularly is. This is shown in Table 1: in the case of
LIME, the focus is on explanations for predictions within a specific part of the data
while in the case of CFE, it is on the input-output relation of an ML model.

17 A similar strategy is pursued by Stuart (forthcoming) who applies the framework of epistemic conse-
nquentialism to the issue of scientific imagination.
Table 1 Possible determination of fine-grained means-end relations for LIME and CFE

| Example | Epistemic end | Means |
|---------|---------------|-------|
|         | What? (Topic) | To whom? (Stakeholder) | How? (Instrument) |
| LIME    | Prediction(s) in specific part of the data | Operators; executors | Local approximation to complex model |
| CFE     | Input-output relation | Decision-subjects | Counterfactual statements |

Answering the question *to whom* something should be explained determines the relevant *stakeholder* who is asking for and receiving the explanation. This is important, since there is a variety of different stakeholders in the ‘ML ecosystem’. The latter term was coined by Tomsett et al. (2018) who distinguish creators (agents that create an ML system), operators (agents interacting directly with an ML system), and executors (agents making decisions informed by an ML system) as well as decision-subjects (agents affected by ML-based decisions) and data-subjects (agents whose personal data has been used to train the system).\(^{18}\) Clearly, this variety of stakeholders suggests a variety of cognitive abilities, background knowledge, or interests across the different agents. It has therefore been pointed out repeatedly that different stakeholders have different explanatory requirements (Langer et al., 2021; Mohseni et al., 2021; Zednik, 2021).\(^{19}\) This is reflected in practice: for instance, LIME is explicitly meant to provide explanations to the users of an ML system while CFE are meant to do the same for data-subjects (Ribeiro, 2016, p. 1135; Wachter, 2018, p. 843).\(^{20}\)

The previous paragraphs reveal that the specification of a topic and the relevant stakeholders further characterizes what aspects should be taken into account when an XAI technique produces an explanation. This determines an epistemic end that is one component of the epistemic means-end relations governing XAI. Clearly then, the corresponding means are the other component. They are determined by answering the question as to *how* an explanation should be achieved, for this specifies the *instruments* deemed appropriate to achieve that explanation at hand. As reflected by the high-level aim in (AIM), this is the main occupation of methodological research in XAI: to develop and improve instruments that produce explanations of ML methods. In the case of LIME, local approximations of complex models are the instruments or means proposed to attain the given epistemic end, while CFE seek to attain the given epistemic end via counterfactual statements (see Table 1).

\(^{18}\) This is only one possible classification of stakeholders. I will use it here as a reference point while acknowledging that, in practice, individuals might be seen as different stakeholders at the same time or as more analytic than the classification suggests (e.g., if a scientist develops and uses an ML model). For another classification, see, for instance, Preece et al. (2018).

\(^{19}\) Recently, it has been argued that, in addition, what matters is the relation between the stakeholder receiving and the stakeholder providing the explanation (Bordt et al., 2022).

\(^{20}\) However, note that there is no standard terminology for the stakeholders in the ML ecosystem yet: the users mentioned by Ribeiro et al. (2016, p. 1135) are likely to be either operators or executors as defined in the classification by Tomsett et al. (2018). Similarly, based on the remarks by Wachter et al. (2018), it seems likely that CFE should rather be provided to decision-subjects rather than to data-subjects as defined in the classification by Tomsett et al. (2018). Thus, to avoid ambiguities, Table 1 was constructed using the latter classification.
Although the exact characterization of LIME and CFE presented in Table 1 might be debatable, the overall framework discussed so far clearly gives rise to a situation that is considerably more fine-grained than the high-level aim of XAI formulated in (AIM): epistemic ends can be split into topic and stakeholder, instruments constitute the corresponding means to achieve given ends. Thus, the high-level aim can be reformulated by stating that XAI seeks to

Provide instruments that produce explanations of topic \( t \) for stakeholder \( s \). (AIM’)

On the one hand, this goes slightly beyond standard means-end epistemology that takes epistemic ends to come in different levels of granularity, yet also takes them as rather monolithic and not separable into further constituents. On the other hand, it hints at a possible strategy to determine the different levels of granularity of epistemic ends in practice: there is a coarse-grained level as expressed in (AIM), stating that an explanation of an ML method should be produced; yet there is a sequence of more fine-grained levels as expressed in (AIM’), at which explanations of specific parts of said ML method should be produced for specific stakeholders. Stating topic, stakeholder, and instrument with increasing precision allows to move from the coarse- to the fine-grained level. As I will show in the following, this kind of analysis has a descriptive and a normative component.

### 3.2 The account’s descriptive component

The means-end account proposed in the previous section has a rather straightforward descriptive component: as illustrated in Table 1, it allows to describe the means-end considerations of XAI researchers by stating what topic and stakeholder constitute their epistemic end and what instruments they propose as means to attain it. As two further examples will show, this is a useful framework for structuring the field of XAI.

First, consider a technique proposed by Hendricks et al. (2016). It is meant to produce explanations of how some visual input to an ML-based image classification system, that is, an image, leads to a specific output, that is, a particular classification. Their solution is to generate textual explanations that describe those parts of the visual input that distinguish it from images in other categories.

Second, consider a technique proposed by Kim et al. (2018). It is meant to produce explanations for the internal state of an ML system that is controlling an autonomous vehicle. Their solution is to generate textual explanations derived from the visual attention of the ML system that controls the vehicle.

Exploiting the means-end account’s descriptive component, it is straightforward to analyze both techniques just like LIME and CFE above. As for the topic, Hendricks et al. (2016) focus on the relationship between an input to the image classification system and its corresponding output. In fact, they make their specification of the topic explicit by stating that they concentrate on “justification explanation systems producing sentences detailing how visual evidence [i.e., some input] is compatible

---

21 In Sect. 4, I will argue that (AIM’) can be refined even further.
with a system output” (Hendricks et al., 2016, p. 3). Furthermore, they contrast their approach with “introspective explanations” that focus on the internal functionality of a system. The latter, however, are the explicit topic defined by Kim et al. (2018, p. 564) who state that they aim at providing “explanations that are based on the system’s internal state”. As for the stakeholder who should receive the explanation, both techniques are meant to address either the operators or the executors of an ML system.

Finally, as for the instruments, Hendricks et al. (2016) propose a technique that consists in describing those parts of input images that are decisive for their classification. To use the authors’ terminology, they propose to generate textual explanations that describe properties that are both “image-relevant” in the sense that the property is really contained in the input at hand and “class-discriminative” in the sense that the property is relevant for distinguishing between different categories (Hendricks et al., 2016, p. 4). For the particular application that the authors investigate, the classification of birds, this leads to explanations like “[t]his is a Western Grebe because this bird has a long white neck, pointy yellow beak and red eye” (Hendricks et al., 2016, p. 2). The explanation mentions the output of the image classification system, ‘Western Grebe’, as well as specific properties that occur in the input image and that distinguish the Western Grebe from other, similar looking birds.

Kim et al. (2018) propose another instrument that is known as visual attention heatmapping to achieve their epistemic end (Kim & Canny, 2017; Xu et al., 2015). This means that their technique to achieve explanations relies on analyzing the visual attention of the ML system controlling the autonomous vehicle. More precisely, the system is confronted with an input of traffic scenes consisting of dashcam videos and sensor measurements such as the vehicle’s speed. The authors employ a so-called attention model to extract salient features of this input that the system ‘looks at’ and that it uses for its decisions regarding the vehicle’s acceleration as well as its change of course. Subsequently, textual descriptions for these salient features are generated, leading to explanations for the vehicle’s behavior such as: “The car is driving forward because there are no other cars in its lane” (Kim et al., 2018, p. 564). Although this explanation resembles the ones produced by the technique of Hendricks et al. (2016), it is generated by a strictly different approach: whereas explanations by Hendricks et al. (2016) are generated conditional on visual properties of the ML system’s input (e.g., ‘long white neck’, ‘pointy yellow beak’, ‘red eye’), explanations by Kim et al. (2018) rely on attention heatmaps and are thus generated conditional on internal states of the ML system. To emphasize this distinction, I refer to the former approach as ‘textual explanations’ and to the latter as ‘verbalized attention heatmapping’.

The new examples just discussed can be used to extend Table 1, giving rise to the situation in Table 2: four XAI techniques are described based on the specific means-end considerations by which they are governed. This is useful for at least two reasons.

---

22 To be precise, both Hendricks et al. (2016) and Kim et al. (2018) mention (end-)users as the relevant stakeholders, which in terms of the classification by Tomsett et al. (2018) are either operators or executors.

23 The whole setup relies on the Berkeley Deep Drive dataset introduced by Xu et al. (2017). Apart from the videos, it contains information about the vehicle’s GPS position and course as well as data retrieved from a gyroscope, a magnetometer, and an inertial measurement unit (Xu et al., 2017, p. 2178).
Table 2  Tentative structure for a taxonomy of XAI techniques

| Example | Epistemic end | Means |
|---------|---------------|-------|
|         | What? (Topic) | To whom? (Stakeholder) | How? (Instrument) |
| …      | …             | …     | …                      |
| LIME    | Prediction(s) in specific part of the data | Operators; executors | Local approximation to complex model |
| CFE     | Input-output relation | Decision-subjects | Counterfactual statements |
| Hendricks et al. (2016) | Input-output relation | Operators; Executors | Textual explanations |
| Kim et al. (2018) | System’s internal state | Operators; Executors | Verbalized attention heatmapping |
| …      | …             | …     | …                      |

First, it reveals that although the XAI literature diverges both conceptually and methodologically, there is a common structure that is shared by the different approaches pursued in the literature. This structure consists in means-end relations that arise from the specification of topics, stakeholders, and instruments.

Second, extending the analysis from above to other XAI techniques paves the way for a taxonomy of the existing literature in the field. Accordingly, it is possible to investigate the means-end relations governing other XAI techniques, to identify the relevant topic, stakeholder, and instrument and to extend Table 2 as indicated by the dots in the first and last row. Thus, the table is not only a description of four XAI techniques, it is also a starting point for a more comprehensive taxonomy of existing methods.

3.3 The account’s normative component

Using the descriptive component of the means-end account, it is possible to unravel the different epistemic ends pursued by different authors and the different means they propose to achieve them. Apparently, authors come up with different instruments when they specify different topics or stakeholders. Thus differences in the ‘what’ and ‘to whom’ of an explanation seem to trigger differences in the ‘how’ of achieving it. This observation might seem entirely unsurprising at first. After all, means-end considerations are common to a variety of contexts different from XAI in which instruments or techniques have to be chosen based on some given end: if I pursue the end of driving a nail into the wall, I better use a hammer instead of a violin bow. Yet if my end consists in practicing my favorite violin sonata, the latter will be much more appropriate. 24 The perspective of means-end epistemology allows to move beyond such intuitively plausible observations. In particular, it helps to make means-

24 Grüne-Yanoff (2021) puts forward a similar argument for the case of scientific methodology.
end relationships in XAI more precise, thereby ultimately leading to the normative component of the present account.

First, recall from (NORM) that means-end epistemology is inherently normative and governed by the principle of instrumental rationality: given an epistemic end, certain means ought to be adopted if and only if they further the given end. Second, we have seen in (FRAME) that problems of XAI can be framed as problems of means-end epistemology. In particular, specifying a topic and stakeholder determines the epistemic end that is pursued, while specifying the corresponding XAI instruments determines the means to attain the given end. Taking these aspects together shows that XAI is inherently normative as well:

\[
\text{Given a topic } t \text{ and stakeholder } s, \text{ certain instruments of XAI ought to be adopted if and only if they further the epistemic end determined by } t \text{ and } s. \quad \text{(NORM}_{XAI})
\]

Put differently, the suitability of XAI techniques depends on what exactly should be explained and to whom the explanation should be given. However, this dependence is a normative one. Indeed, differences in the topic and stakeholder determining an epistemic end trigger differences in the instruments to achieve it, but they do so with normative force.

For instance, it is not by mere coincidence that the last column of Table 2 displays four different instruments. Rather, the topics and stakeholders defined by the different authors determine different epistemic ends. The means-end account’s normative component tells us that different ends call for different means. So clearly, given different topics or stakeholders, the specific instrument proposed by Hendricks et al. (2016) differs from the one proposed by Kim et al. (2018) which in turn differs from CFE. Furthermore, Table 2 highlights the importance of analyzing epistemic ends at a high level of granularity. For instance, both the technique proposed by Kim et al. (2018) and the one proposed by Hendricks et al. (2016) are meant to provide explanations to either operators or executors of an ML system, yet the former authors aim at explanations of a system’s internal states whereas the latter aim at explanations of its input-output relation. Thus, while addressing the same stakeholders, the techniques are meant to produce explanations of different topics, that is, of different aspects of an ML system. In sum, this leads to different epistemic ends being pursued and different means ought to be adopted to achieve them.

Consequently, the examples hint at what the normative component allows to achieve on a more general level: it allows to spell out in great detail what an explanation for an ML method should ‘look like’ in a specific setting. Thus, analyzing particular ends, that is, particular topics and stakeholders, can reveal which instruments ought to be

---

25 To be precise, it tells us that different ends may call for different \emph{sets} of means, \( M \) and \( M' \). The latter can be partially overlapping, such that \( M \cap M' \neq \emptyset \), and they become more specific as the description of the end becomes more fine-grained (see Sect. 2.2).

26 The same reasoning could be applied when comparing the technique proposed by Hendricks et al. (2016) to CFE. In that case, the topic is the same, but explanations should be produced for different stakeholders which, in sum, also leads to different epistemic ends (see Table 2).
adopted to achieve them and which instruments can be ruled out as inappropriate. In order to allow for such insights, it is important to recognize the logical form in which the principle of instrumental rationality (NORM) is stated in means-end epistemology and to preserve this form also in (NORM_XAI). Topic and stakeholder constitute the epistemic end which is in turn tied by a biconditional to the adoption of certain means, that is, certain instruments of XAI: if certain instruments ought to be adopted, then they further the epistemic end determined by the given topic and stakeholder; if they further the epistemic end determined by the given topic and stakeholder, then certain instruments ought to be adopted. Considering only one of the latter conditional statements could give rise to two types of problematic cases: on the one hand, cases in which certain instruments ought not to be adopted, although they further the given epistemic end. On the other hand, cases in which certain instruments do not further the given epistemic end, yet nevertheless ought to be adopted.

However, if such problematic cases are prevented, the normative component can be used to show how the set of appropriate XAI instruments is constrained by the particular ends for which an explanation is sought. As outlined above, in means-end epistemology and in the means-end account of XAI alike, the set of appropriate instruments gets narrower as the epistemic end gets more precise. This implies that one should aim for maximally specific topics and stakeholders to identify a maximally specific set of appropriate instruments that one ought to adopt. Furthermore, the central insight of means-end epistemology that “all we need to provide a means-end analysis […] is a sufficiently clear description of the goals in question” (Schulte, 1999, p. 26) hints at the precondition for assessing the suitability of XAI techniques in a given context: we can only ever determine whether an XAI technique is appropriate, if ex ante, the epistemic end being pursued is specified in sufficient detail. This implication of the present investigation is so far largely neglected in the discourse on XAI. In particular, high-level regulation such as the EU General Data Protection Regulation or the European Union’s draft Artificial Intelligence Act does not spell out the precise epistemic end corresponding to the transparency requirements imposed on ML methods.

Thus overall, the means-end account’s normative component establishes a normative framework for the field of XAI. The framework is based on the insight that a close analysis of what specific part of an ML method should be explained and to whom it should be explained allows one to identify the set of instruments that is appropriate

---

27 For two examples in the latter vein, see Sect. 4.
28 There might be cases in which certain instruments cannot be adopted, although they further the given epistemic end, for instance due to privacy issues or proprietary software. Yet this does not affect the normative result that, in principle, they nevertheless ought to be adopted.
29 Note that no matter how specific topic and stakeholder are defined, there will be several appropriate instruments and, hence, a set of different particular means for the same end in most cases (Mothilal et al., 2021).
30 As mentioned in the outline of means-end epistemology above, this is the reason why it seems preferable, if there are several epistemic ends, to analyze them separately rather than as a set: the hypothetical part of (NORM), ‘given an epistemic end’, gets more precise, which allows us to formulate a more precise imperative as well (‘these particular means ought to be adopted’).
to bring about that specific explanation. This unifies and extends existing proposals that point to a similar direction.

First, the means-end account’s normative component unifies rather descriptive frameworks of XAI like the one proposed by Sokol & Flach (2020). They carve out “a set of descriptors that can be used to characterise and systematically assess explainable systems” (Sokol & Flach, 2020, p. 56). However, although leading to a useful and very detailed taxonomy of XAI techniques, the framework does not provide an overarching account of how the assessment of XAI techniques connects them to specific situations. This can be achieved using the means-end account’s normative component: a specific situation will be characterized by a specific epistemic end being pursued which in turn normatively prescribes a set of appropriate instruments. For instance, consider a loan applicant—who asks for an explanation of why their loan application was accepted or rejected by the bank’s ML-based decision system. The applicant’s explanatory topic—the decision, that is, the output of the ML-based system—and their own role as a specific stakeholder with specific explanatory requirements that need to be fulfilled constitute an epistemic end that clearly rules out a highly specific explanation of the system’s internal processes as the appropriate means. Instruments that provide some high-level explanation of why the system’s input led to a specific output might be more appropriate in this case. Once such a set of appropriate instruments has been established, a purely descriptive account can help to assess whether a given XAI technique belongs to that set or not. Yet beforehand, it is epistemic normativity that links the characteristics of a situation to a set of instruments that are appropriate due to their specific characteristics and that hence ought to be adopted.

Second, the means-end account’s normative component goes beyond other proposals of normative frameworks that solely focus on the explanatory requirements of different stakeholders. For instance, on Zednik’s (2021) account, these requirements can be characterized by spelling out the epistemically relevant elements of an ML system for a given stakeholder. Since these elements differ across stakeholders, different explanatory requirements have to be satisfied by tailor-made explanations. Thus, an XAI technique should be used if it produces explanations that fulfill a stakeholder’s explanatory requirements in virtue of addressing the specific epistemically relevant elements. However, fulfilling the requirements of different stakeholders can hardly be all there is to the suitability of XAI techniques, since one can easily imagine a situation in which the same stakeholder asks for different explanations. For instance, consider once more a loan applicant, but this time with a genuine interest in ML. In that case, they might on the one hand still ask for an explanation of why their loan application was accepted or rejected, that is, why certain inputs led to a certain output. On the other hand, however, they might also ask for an explanation of the ML system’s internal functioning. As already shown in Table 2 above, the means-end account of XAI accommodates both situations as in this example and accounts that derive different explanatory requirements from differences across stakeholders in a straightforward way.

31 The deliberately open formulation ‘allows one to identify’ is meant to leave room for different interpretations: in the general setting of means-end epistemology, it refers to the epistemologist analyzing means-end relations; in practice, it can refer to developers of XAI techniques, but similarly to regulators trying to substantiate the requirements of high-level regulation.
way: on this account, the explanatory requirements of different stakeholders are but one source of variation in the epistemic end of XAI, another one consisting in the specification of different topics. The means-end account’s descriptive component allows to distinguish them accurately and to analyze their particular specifications as well as the epistemic end that they determine. Additionally, given a certain epistemic end, the account’s normative component allows to identify the set of instruments that the end calls for.

4 Extending the account

We have made quite some progress up to this point: we have seen that the high-level aim of XAI can be reformulated in a more granular way (see (AIM')), that problems of XAI can be framed as problems of means-end epistemology (see (FRAME)), and that, consequently, XAI is inherently normative (see (NORMXAI)). However, the preceding discussion exclusively focused on epistemic aspects, considering the epistemic end of producing explanations to be the main ingredient of the means-end account. Albeit leading to an account that accurately reflects the structure of XAI problems and is useful for their analysis, this perspective needs to be broadened for at least two reasons.

First, there can be epistemic ends other than that of producing explanations. Indeed, it has been pointed out that often, XAI techniques should produce explanations to achieve further epistemic ends such as understanding or interpretability (Erasmus et al., 2021). Seen this way, achieving the epistemic end of producing explanations can also turn into a means to achieve another epistemic end.32

Second, XAI techniques are commonly employed to achieve non-epistemic ends. For instance, returning to the examples from above, LIME and the technique proposed by Hendricks et al. (2016) are meant to foster trust in an ML system; Kim et al. (2018) aim at enabling human agents to extrapolate the behavior of the vehicle controlled by an ML system; CFE are meant to inform, provide grounds to contest adverse decisions, and reveal how to reach a desired result in the future (Wachter et al., 2018, p. 843). Clearly, several of the latter ends are not or at least not strictly confined to the epistemic realm.

Thus, beyond the epistemic ends discussed above, there are further ends relevant to XAI that might stem from epistemic or non-epistemic considerations. So in addition to answering what should be explained, to whom it should be explained, and how it should be explained, researchers seem to answer the question as to why something should be explained to begin with. I call the answer to this question the goal of an explanation. Incorporating this into the means-end account, one can extend (AIM') by stating that XAI seeks to

32 This parallels more general philosophical arguments that explanations of a phenomenon are not (only) intrinsically valuable, but (also) instrumentally valuable because they lead to understanding of the phenomenon (Grimm, 2010).
Provide instruments that produce explanations of topic $t$ for stakeholder $s$ to achieve goal $g$. (AIM*)

This extension leads to a setup in which there are two types of means-end relations: the strictly epistemic means-end relations discussed above and means-end relations that can, but need not be, strictly epistemic, depending on the specification of the goal. Importantly, the former relations are instrumental to the latter ones, since achieving the epistemic end of producing explanations of topic $t$ for stakeholder $s$ can also be a means to achieving goal $g$, provided that the particular epistemic end really is suitable to achieve the particular goal. This setup could easily be extended even further to what one might call a cascade of means-end relations in which achieving the end of one relation is also a means to achieve the end of the next relation—and so on. To identify a suitable XAI technique, one would then have to work backwards through the cascade, ultimately arriving at one of the strictly epistemic means-end relations discussed above. Extending the means-end account in this way is important, since it reflects both the methodological and the conceptual literature on XAI.

As for the methodological literature, consider once more the two exemplary XAI techniques introduced at the beginning, LIME and CFE. For CFE, I just pointed out that, among other things, they are meant to reveal “what could be changed to receive a desired result in the future” (Wachter et al., 2018, p. 843), a problem that is commonly referred to as algorithmic recourse (Venkatasubramanian & Alfano, 2020). Based on the framework developed in this paper, we can assess whether CFE live up to this goal. The statement in (AIM*) tells us that, to achieve it, CFE first have to achieve a subordinate epistemic end that is instrumental to the goal, namely, to produce a suitable explanation of some topic for some stakeholder. Using the descriptive component of the means-end account, we could already identify the epistemic end that is pursued by CFE: as shown in Table 2, it consists of a topic corresponding to an ML model’s input-output relation and of decision-subjects as the relevant stakeholder. The crucial questions then are whether CFE are appropriate to achieve this epistemic end and, if so, whether this epistemic end is instrumental to the goal of algorithmic recourse. Providing information about inputs and outputs in a way that does not presuppose any familiarity with ML, the counterfactual statements produced by CFE are indeed an appropriate means to achieve the epistemic end of producing explanations of an ML model’s input-output relation for a decision-subject. Thus, according to (NORMXAI), CFE ought to be adopted given this epistemic end. However, by exclusively focusing on the input-output relation of the ML model, CFE ignore “the causal relationships governing the world in which actions will be performed” (Karimi et al., 2021, p. 353). This means that although CFE recommend a set of alternative actions (e.g., ‘increase annual income by amount X’) these need not lead to the desired result (e.g., ‘you were granted the loan’), since the mechanism by which the ML model operates might not properly reflect the true underlying mechanism. Consequently, while the epistemic end pursued by CFE is an appropriate means to achieve the goal of identifying and contesting adverse decisions, it is inappropriate to achieve the goal.

33 Apart from being intuitively plausible, this strategy of working backwards is suggested by the literature on means-end reasoning in both philosophy and artificial intelligence (Bratman, 1981; Pollock, 1998).

34 An intuitive example for this situation is provided in Karimi et al. (2021, p. 353).
of providing information on how to turn them into a desired result in the future. The preceding remarks show that given the latter goal, a topic that is different from an ML model’s input-output relation should be specified—and, hence, that CFE ought not to be used in this case, because they would be inappropriate to achieve this different epistemic end.

With the means-end account of XAI at hand, a similar analysis can be conducted for LIME: as shown in Table 2, this technique seeks to achieve the epistemic end of explaining predictions in a specific part of the data to the operators or executors of an ML model by locally approximating the model. Furthermore, achieving this epistemic end should ultimately achieve the goal of increasing the users’ trust in the ML model. Similar to the case of CFE, the crucial questions then are whether LIME is appropriate to achieve the given epistemic end and, if so, whether this epistemic end is instrumental to the goal of increasing trust. Yet different from the case of CFE where the specification of the topic turned out to be problematic, we should now have a closer look at the specification of the relevant stakeholder to answer these questions. Recall that, according to Ribeiro et al. (2016), explanations produced by LIME are targeted at the users of an ML model which correspond to either operators or executors in the classification by Tomsett et al. (2018). Thus, either way, the stakeholders specified as part of the epistemic end pursued by LIME are crucially distinct from the creators of the ML model. They might therefore be able to use the ML model by making some clicks in a user interface, but nevertheless lack an adequate background knowledge in mathematics let alone ML that would render the local approximation produced by LIME useful for them. Accordingly, it seems doubtful whether LIME really is an appropriate instrument to achieve the specified epistemic end and, hence, whether it ought to be adopted. In line with (AIM*), this doubt also carries over to the goal of increasing trust: at first, it seems intuitive to assume that explaining the predictions of an ML model to its users increases the trust they place in the model.35 However, this presupposes that the means producing the explanation are appropriate for the relevant stakeholders which does not seem to be warranted in the case of LIME.36 In fact, to achieve the goal of increasing the trust of users in an ML model, it may be beneficial to adjust the epistemic end, in particular the topic, and to adopt an instrument like the one proposed by Hendricks et al. (2016) that produces a simple textual explanation and, hence, does not require any background knowledge in ML whatsoever. Apart from the importance of rigorous means-end considerations in the context of XAI, this example also reveals the importance of the epistemic conception of explanation introduced above (p. 5): what counts as a proper explanation in the context of XAI should mainly depend on epistemic considerations, for instance, on whether stakeholders possess the necessary background knowledge to benefit from the explanations produced by LIME or whether they find a textual explanation more intuitive.

As for the conceptual literature, recall from above that Krishnan (2020, p. 488) argues that what is called explainability or interpretability “serves as a means to an

35 This is discussed as the explainability-trust hypothesis (Kästner et al., 2021).
36 Indeed, Ribeiro (2016, p. 1143) acknowledge that to validate whether LIME increases trust, they relied on a group of “graduate students who have taken at least one graduate machine learning course”. This has also been observed by Blanco (2022, p. 249).
end, rather than being an end in itself’’ (Krishnan 2020, p. 488). Thus, beyond the basic epistemic end of producing explanations and beyond further epistemic ends such as interpretability, there are more ‘fundamental goals’ (Krishnan 2020, p. 495) that she takes to be the actual ends that should be achieved. Accordingly, problems of XAI should be framed in a way that uncovers these fundamental goals to “facilitate a more pluralistic approach to problem-solving” (Krishnan 2020, p. 495). So on her view, identifying XAI techniques that are appropriate in a particular situation requires a fine-grained characterization of that situation in the first place. The extension introduced in (AIM*), distinguishing epistemic ends from further goals that might be specified, offers a straightforward way to accommodate this line of argumentation within the means-end account of XAI. First, we have seen that epistemic ends come in varying degrees of granularity. Second, the account’s normative component reveals that more or less specific ends entail more or less specific sets of appropriate instruments. Third, different (sets of) means ought to be adopted to achieve different epistemic ends. Fourth, different epistemic ends are required to achieve different goals. Consequently, the means-end account of XAI allows for a fine-grained investigation of what Krishnan refers to as ‘fundamental goals’. It also reveals why a pluralistic approach to problem-solving is indeed necessary in XAI: since one ought to adopt the means to one’s ends.

Having said all this, however, one might object that the perspective of normative epistemology adopted in this article is no longer essential to the means-end account. After all, what matters when faced with a cascade of means-end relations seems to be the strategy of backward-planning mentioned above, since this suffices to ensure means-end coherence across the different relations. But we do not have to consider an entire cascade of means-end relations, we do not even have to go beyond statement (AIM*) to notice the following: XAI is an epistemic endeavor at heart. No matter what further goals are specified, we have seen that a strictly epistemic means-end relation is the nucleus of everything that follows. This is why the perspective of normative epistemology is crucial.

5 Conclusion

This paper set out by observing considerable disagreement in the XAI literature, both on a conceptual and on a methodological level. Does this mean that the field of XAI is scattered into various disconnected subprojects? Quite to the contrary. Indeed, I argue that there is a common structure that is shared by the variety of approaches pursued in the literature. To do so, I put forward a means-end account of XAI. The account relies on the insight from means-end epistemology that one ought to adopt appropriate means to one’s epistemic ends. It also relies on the observation that problems of XAI can be framed as problems of means-end epistemology. Taking both aspects together, I show that normative means-end relations are and should in fact be central to XAI. I further show that these means-end relations are determined by a topic, a stakeholder, a goal, and an instrument. By specifying these, one provides answers to four central

---

37 Similar arguments have been put forward, e.g., by Bordt et al. (2022).
questions: What should be explained? To whom should it be explained? Why should it be explained? And how should it be explained?

The means-end account of XAI has several important consequences. First, it explains why disagreement arises in the field: the divergence in instruments of XAI follows from the disagreement on epistemic ends. Second, it structures the field: there is a common methodology of developing appropriate means for given ends. Third, this structure has a descriptive component: different authors specify different ends and come up with different means to achieve them. This gives rise to a taxonomy that classifies existing contributions to the field along the specific means-end relations that are considered. Fourth, this structure also has a normative component: the ends of an explanation normatively constrain the set of admissible means to achieve it. The means-end account thus reveals how the suitability of particular instruments of XAI is prescribed by the ends for which an explanation is sought.

Future research might investigate the different components of the means-end account even further. On the one hand, I plan to establish a more comprehensive taxonomy of existing XAI techniques using the account’s descriptive component and to evaluate it using the normative component. On the other hand, there is potential to apply the means-end account to regulatory issues. From this perspective, the account provides all ingredients that are needed for aligning the right XAI techniques with the right situations and thus, ultimately, for designing effective guidelines of XAI.

Acknowledgements I would like to thank Sara Mann for helpful conversations and audiences in Tübingen as well as in Frankfurt a. M., in particular David Danks, Roman Heil, Kate Vredenburgh, and Alexandra Zinke for helpful comments. I would also like to thank Sara Blanco, Karoline Reinhardt as well as Thomas Grote for helpful feedback and Wolfgang Spohn for very valuable comments on an earlier draft. Finally, I would like to thank Eric Raidl for his guidance and continuing support during the entire process that led up to this article.

Funding Open Access funding enabled and organized by Projekt DEAL. My research was funded by the Baden-Württemberg Foundation (program “Verantwortliche Künstliche Intelligenz”) as part of the project AITE (Artificial Intelligence, Trustworthiness and Explainability).

Declarations

Competing interest There are no competing interests to declare.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Ahlstrom-Vij, K., & Dunn, J. (2014). A defence of epistemic consequentialism. The Philosophical Quarterly, 64(257), 541–551.
Baldi, P., Sadowski, P., & Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. *Nature Communications, 5*(4308), 1–9.

Besold, T., & Uckelmann, S. (2018). The what, the why, and the how of artificial explanations in automated decision-making. https://arxiv.org/abs/1808.07074.

Blanco, S. (2022). Trust and explainable AI: Promises and limitations. In Koskinen, J., Kimppa, K. K., Heimo, O., Naskali, J., Ponkala, S., & Rantanen, M. M., (eds.), *Proceedings of the ETHICOMP 2022* (pp. 246–257). Turku: University of Turku.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., & Zieba, K. (2016). End to end learning for self-driving cars. https://arxiv.org/abs/1604.07316.

Bordt, S., Finck, M., Raidl, E., & von Luxburg, U. (2022). Post-hoc explanations fail to achieve their purpose in adversarial contexts. In *FAccT ’22: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency* (pp. 891–905).

Bratman, M. (1981). Intention and means-end reasoning. *The Philosophical Review, 90*(2), 252–265.

Burrell, J. (2016). How the machine ‘Thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science, 87*(4), 568–589.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118.

Finlay, S. (2009). Oughts and ends. *Philosophical Studies, 143*(3), 315–340.

Foreman-Mackey, D., Montet, B. T., Hogg, D. W., Morton, T. D., Wang, D., & Schölkopf, B. (2015). A systematic search for transiting planets in the K2 data. *The Astrophysical Journal, 806*(2), 215–228.

Garreau, D., & von Luxburg, U. (2020). Explaining the explainer: A first theoretical analysis of LIME. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, PMLR, 108*, 1287–1296.

Grimm, S. R. (2010). The goal of explanation. *Studies in History and Philosophy of Science Part A, 41*(4), 337–344.

Keil, F. C. (2006). *Belief and counterfactuals. A study in means-end philosophy*. New York: Oxford University Press.

Joyce, J. M. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science, 65*(4), 575–603.

Karimi, A.-H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions. In *FAccT ’21: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 353–362).

Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., & Sterz, S. (2021). On the relation of trust and explainability: Why to engineer for trustworthiness. 2021 *IEEE 29th international requirements engineering conference workshops (REW)* (pp. 169–175).

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology, 57*(1), 227–254.

Kim, J., & Canny, J. (2017). Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision* (pp. 2942–2950).

Kim, J., Rohrbach, A., Darrell, T., Canny, J., & Akata, Z. (2018). Textual explanations for self-driving vehicles. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer vision—ECCV 2018* (pp. 563–578). Cham: Springer.
Krishnan, M. (2020). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology, 33*(3), 487–502.

Langer, M., Oster, D., Speith, T., Hermanns, H., Küstner, 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. *Artificial Intelligence, 296*, 103473.

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue, 16*(3), 31–57.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10*(10), 464–470.

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. In *NIPS’17: proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777).

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

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems, 11*(3–4), 1–45.

Mothilal, R. K., Mahajan, D., Tan, C., & Sharma, A. (2021). Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *AIES ‘21: Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society* (pp. 652–663).

Nyup, R., & Robinson, D. (2022). Explanatory pragmatism: A context-sensitive framework for explainable medical AI. *Ethics and Information Technology, 24*(1), 1–15.

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

Pollock, J. L. (1998). The logical foundations of goal-regression planning in autonomous agents. *Artificial Intelligence, 106*(2), 267–334.

Potochnik, A. (2016). Scientific explanation: Putting communication first. *Philosophy of Science, 83*(5), 721–732.

Preece, A., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). Stakeholders in explainable AI. https://arxiv.org/abs/1810.00184.

Raidl, E., & Rott, H. (forthcoming). Towards a logic for ‘Because’. *Philosophical Studies*. https://doi.org/10.1007/s11098-023-01998-4.

Räz, T. (2022). Understanding deep learning with statistical relevance. *Philosophy of Science, 89*(1), 20–41.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

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

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton, N.J.: Princeton University Press.

Salmon, W. C. (1989). Four decades of scientific explanation. In P. Kitcher & W. C. Salmon (Eds.), *Scientific explanation, Minnesota studies in the philosophy of science* (pp. 3–219). Minneapolis: University of Minnesota Press.

Schubbach, A. (2021). Judging machines: Philosophical aspects of deep learning. *Synthese, 198*(2), 1807–1827.

Schulte, O. (1999). Means-ends epistemology. *The British Journal for the Philosophy of Science, 50*(1), 1–31.

Sokol, G., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *FAT* ‘20: *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 56–67).

Stuart, M. T. (forthcoming). Scientists are epistemic consequentialists about imagination. *Philosophy of Science*. https://doi.org/10.1017/psa.2022.31.

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

Venkatasubramanian, S., & Alfano, M. (2020). The philosophical basis of algorithmic recourse. In *FAT* ‘20: *Proceedings of 2020 ACM conference on fairness, accountability, and transparency* (pp. 284–293).
Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.

Woodward, J., & Ross, L. (2021). Scientific explanation. In Zalta, E. N. (eds.), *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/.

Xu, H., Gao, Y., Yu, F., & Darrell, T. (2017). End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2174–2182).

Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd international conference on machine learning* (pp. 2048–2057).

Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology, 34*(2), 265–288.

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology, 32*(4), 661–683.

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