A Means-End Account of Explainable Artificial Intelligence

*Oliver Buchholz**

August 10, 2022

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.

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). 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.[2]

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 and Kim 2017, Erasmus et al. 2020), the more intuitive *understandability* (Páez 2019) as well as the larger category of *transparency* (Günther and 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 and 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

---

[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] 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 inherently interpretable methods, thereby avoiding the use of opaque ML methods from the outset (Rudin 2019).

insights from means-end epistemology.[3] *Means-end epistemology* takes epistemology to be a normative discipline; it is based on the principle of instrumental rationality (Huber 2021: 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 2021: 125, Schulte 1999: 7). Providing an explanation is such an epistemic end, because it relates to the epistemic state of the agent receiving the explanation.[4] 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.[5] Indeed, I show that pragmatic considerations are relevant, especially in what concerns the topics, stakeholders,

---

[3]The idea of exploiting existing research from philosophy and the social sciences when analyzing XAI is put forward, e.g., in Miller (2019).

[4]This is also a well-established result in psychology as reported, for instance, by Keil (2006) or Lombrozo (2006).

[5]Examples include Besold and Uckelmann (2018), Nyrup and Robinson (2022), and Páez (2019).

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.[6]

The remainder of this article is structured as follows: In Section 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 Section 3, I argue that problems of XAI can be framed as problems of means-end epistemology. I thereby establish the means-end account of XAI and discuss both its normative and its descriptive component. In Section 4, I discuss an extension of the account before concluding in Section 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

XAI is commonly employed to eliminate the inherent opacity of many ML methods and models. In a first step, there is usually some opaque ML model, for instance a deep neural network, whose opacity should be (at least partly) eliminated. In a second step, some XAI technique is applied to the opaque ML model to achieve that elimination. 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. To do so, LIME approximates the possibly highly complex ML model by a simpler model around the specific prediction(s) of interest. Thus, a complex decision surface might be approximated locally by a linear function. This makes the overall ML model locally *interpretable*, thereby revealing the model's structure within a specific part of the data. Since LIME is *model-agnostic*, it is applicable to a wide range of ML methods.

Second, consider counterfactual explanations (CFE) (Wachter et al. 2018). They are usually provided as verbalized counterfactual statements that take the following form:

---

[6]Other normative frameworks for XAI have been proposed, for instance, by Langer et al. (2021), Mohseni et al. (2021), and Zednik (2019).

"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.[7] As becomes obvious, CFE exclusively focus on a model's input-output relation: they explain why a certain input led to a certain output "without opening the black box", that is, without considering a model's internal functioning. They justify the model's output by giving reasons rather than by detailing the decision rule that led to the output.

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

(1)   Provide instruments that produce explanations of ML methods.

Clearly, the term 'explanations' in (1) is dangerously loaded, both within XAI and within philosophy. In this article, I use it 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 definition or not. At this stage, my sole aim is to ensure a maximally general formulation of the high-level goal of XAI that is meant to be as uncontroversial as possible. I will return to it below when putting forward the main argument.

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):

---

[7]I use the rather cautious 'seem', since this view is controversial. See Section 3.3 for a brief overview about the debate in XAI. See also Raidl and Rott (2022) as well as references therein for the general philosophical debate about the relation of counterfactuals and 'because'-statements.

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

Thus, agents are assumed to have certain epistemic ends and the principle of instrumental rationality requires them to adopt those means that are appropriate to achieve the given end(s).[8] For instance, consider Anne, who pursues the epistemic end of holding only true beliefs. She falsely believes that she was denied a loan because her annual salary is too low. Now suppose that she receives the (true) information that, in fact, she was denied the loan because her loan application contained transposed digits for her annual salary. According to means-end epistemology, Anne—if rational—ought to adopt means that lead to a revision of her initial and false belief in favor of the true belief that her loan application contained transposed digits for her annual salary, since these means would further her epistemic end of holding only true beliefs.

Equating epistemic normativity with instrumental rationality and expressing it in terms of means-end relationships as in (2) involves several aspects that are important for the subsequent discussion. First, different ends call for different means. Thus, 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: 26). This is a consequence of the normative perspective adopted by means-end epistemology.

Second, note that 'certain means' in (2) 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, Schulte (1999) conducts a means-end analysis of inductive inference. His starting point 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 and 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.

Fourth, note that the means-end statement in (2) has the logical form of a biconditional and is therefore equivalent to a conjunction of two conditional statements pointing in two opposite 'directions'[9]: according to one direction, *if* certain means ought to be adopted,

---

[8]Note, that (2) expresses a *hypothetical* rather than a *categorical* imperative. It is controversial whether 'ought' needs to be replaced by 'should' in this case (Finlay 2009, Foot 1972, Huber 2021: Ch. 5.4). Since engaging with this debate is beyond the scope of this text, my use of 'ought' is only meant to emphasize the pronounced normative perspective of means-end epistemology.

[9]This is due to the elementary fact that for two propositions P and Q, the biconditional $P \leftrightarrow Q$ holds

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 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 (1): XAI seeks to provide instruments that produce explanations of ML methods. On the one hand, this means that XAI clearly seeks to affect the epistemic states of individuals, 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 by producing explanations, XAI pursues an epistemic end. On the other hand, XAI seeks to provide methods to achieve this epistemic end. Put differently, XAI seeks to find appropriate means to further a given epistemic end. Taken together, these aspects reveal that

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

One might object that it is somewhat circular to argue for (3) based on the high-level goal of XAI proposed in (1), since the latter could have been purposefully designed for precisely this argument. However, as mentioned above, (1) is simply formulated on the highest level possible so as to give rise to a maximally broad and uncontroversial statement.

Let me illustrate the insight in (3) using the examples from Section 2.1. 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.

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

----

if and only if the conjunction of the single conditionals $P \rightarrow Q$ and $Q \rightarrow P$ holds.

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.

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).[10] Clearly, this variety of stakeholders suggests a variety of cognitive abilities, previous 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 2019).[11] 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 et al. 2016: 1135, Wachter et al. 2018: 843).[12]

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 corre-

---

[10]This is only one possible classification of stakeholders and I will use it here as a reference point. For another classification, see, for instance, Preece et al. (2018).

[11]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).

[12]However, note that there is no standard terminology for the stakeholders in the ML ecosystem yet: the users mentioned by Ribeiro et al. (2016: 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-* 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.

Table 1. Possible determination of fine-grained means-end relations for LIME and CFE.

	Epistemic End		Means
Example	*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

sponding 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 (1), 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).

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 (1): 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

(1')   Provide instruments that produce explanations of topic t for stakeholder s.

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 (1), stating that an explanation of an ML method should be produced; yet there is a sequence of more fine-grained levels as expressed in (1'), 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 with a system output" (Hendricks et al. 2016: 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: 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.[13]

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: 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: 2). The explanation mentions the

---

[13]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.

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
. . .	. . .	. . .	. . .

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 and 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 continuously confronted with an input of traffic scenes consisting of dashcam images 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: 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.

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 allows 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.[14] The perspective of means-end epistemology allows to move beyond such intuitively plausible observations. In particular, it helps to make means-end relationships in XAI more precise, thereby ultimately leading to the normative component of the present account.

First, recall from (2) 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 (3) 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:

(4)   Given a topic t and stakeholder s, certain instruments of XAI ought to be adopted if and only if they further the epistemic end determined by t and s.

Put differently, the suitability of XAI techniques depends on what exactly should be

---

[14]Grüne-Yanoff (2021) puts forward a similar argument for the case of scientific methodology.

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.[15]

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 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 (2) is stated in means-end epistemology and to preserve this form also in (4). Topic and stakeholder constitute the epistemic end which is in turn tied by a biconditional to the adoption of certain means, that is, certain intruments 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.[16] On the other hand, cases in which

---

[15]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).

[16]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

13

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.[17] 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: 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 to determine the set of instruments that is appropriate 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 and Flach (2020). They carve out "a set of descriptors that can be used to characterise and systematically assess explainable systems" (Sokol and Flach 2020: 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—without expertise in ML—who asks for an explanation of why her loan application was accepted or rejected by the bank's ML-based decision system. The applicant's explanatory topic—

---

the normative result that, in principle, they nevertheless *ought* to be adopted.

[17]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 to the same end in most cases (Mothilal et al. 2021).

the decision, that is, the output of the ML-based system—and her 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 (2019) 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, she might on the one hand still ask for an explanation of why her loan application was accepted or rejected, that is, why certain inputs led to a certain output. On the other hand, however, she 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 accomodates both situations as in this example and accounts that derive different explanatory requirements from differences across stakeholders in a straightforward 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 (1')), that problems of XAI can be framed as problems of means-end epistemology (see (3)), and that, consequently, XAI is inherently normative (see (4)). 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 sometimes, XAI techniques should produce explanations to achieve further epistemic ends such as understanding or interpretability (Erasmus et al. 2020). Seen this way, achieving the epistemic end of producing explanations can also turn into a means to achieve another epistemic end.

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: 843).

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. I call the answer to this question the *goal* of an explanation. Incorporating this into the means-end account, one can extend (1') by stating that XAI seeks to

(1*)   Provide instruments that produce explanations of topic t for stakeholder s to achieve goal g.

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 determine 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.[18]

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 example of CFE. As discussed above, CFE are meant to reveal "what could be changed to receive a desired result in the future" (Wachter et al. 2018: 843), a problem that is commonly referred to as *algorithmic recourse* (Venkatasubramanian and Alfano 2020). However, it has been claimed that CFE fall short of achieving this goal, for they ignore "the *causal relationships* governing the world in which actions will be performed" (Karimi et al. 2021: 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 reflect the true underlying mechanism properly.[19] The debate surrounding CFE reveals that, although implicitly, debates about the suitability of XAI methods indeed center around means-end considerations and their inherent normativity: according to the literature, counterfactual statements are an appropriate means to achieve the epistemic end of producing explanations of an ML system's input-output relation for a decision-subject (see Table 1). Furthermore, this epistemic end seems to be an appropriate means to achieve the goal of identifying and contesting adverse decisions. However, it seems to be an inappropriate means to achieve the goal of providing information on how to turn them into a desired result in the future. Consequently, CFE ought not be used when the latter is the goal.[20]

As for the conceptual literature, recall from above that Krishnan (2020: 488) argues that what is called explainability or interpretability "serves as a means to an end, rather than being an end in itself" (Krishnan 2020: 488).[21] Thus, beyond the basic epistemic end of producing explanations and beyond further epistemic ends such as interpretability, there are more 'fundamental goals' (Krishnan 2020: 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: 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 (1*), distinguishing epistemic

---

[18]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).

[19]An intuitive example for this situation is provided in Karimi et al. (2021: 353).

[20]Similarly, it has been questioned whether producing particular explanations is an appropriate means to achieve the goal of establishing human trust in an ML system (Kästner et al. 2021).

[21]Similar arguments have been put forward, e.g., by Bordt et al. (2022).

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 (1*) 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 XAI can be framed as a problem 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 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 paves the way for designing effective guidelines of XAI that ensure that the right techniques are employed in the right situations.

References

Baldi, P., Sadowski, P., and Whiteson, D. (2014). Searching for Exotic Particles in High-Energy Physics with Deep Learning. *Nature Communications*, 5(4308):1–9.

Besold, T. and Uckelmann, S. (2018). The What, the Why, and the How of Artificial Explanations in Automated Decision-Making. *arXiv*, 1808.07074:1–20.

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., and Zieba, K. (2016). End to End Learning for Self-Driving Cars. *arXiv*, 1604.07316.

Bordt, S., Finck, M., Raidl, E., and von Luxburg, U. (2022). Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, ACM Digital Library, pages 891–905, New York, NY. Association for Computing Machinery.

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. and Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv*, 1702.08608.

Erasmus, A., Brunet, T. D. P., and Fisher, E. (2020). What is Interpretability? *Philosophy & Technology*, 174(2):1–30.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and 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.

Foot, P. (1972). Morality as a System of Hypothetical Imperatives. *The Philosophical Review*, 81(3):305.

Foreman-Mackey, D., Montet, B. T., Hogg, D. W., Morton, T. D., Wang, D., and Schölkopf, B. (2015). A Systematic Search for Transiting Planets in the K2 Data. *The Astrophysical Journal*, 806(2):215–228.

Grüne-Yanoff, T. (2021). Justifying Method Choice: A Heuristic-Instrumentalist Account of Scientific Methodology. *Synthese*, 199(1-2):3903–3921.

Günther, M. and Kasirzadeh, A. (2021). Algorithmic and Human Decision Making: For a Double Standard of Transparency. *AI & SOCIETY*.

Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016). Generating Visual Explanations. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 3–19, Cham. Springer International Publishing.

Huber, F. (2021). *Belief and Counterfactuals. A Study in Means-End Philosophy*. Oxford University Press, New York.

Karimi, A.-H., Schölkopf, B., and Valera, I. (2021). Algorithmic Recourse: from Counterfactual Explanations to Interventions. *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362.

Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T., and Sterz, S. (2021). On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pages 169–175.

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

Kim, J. and Canny, J. (2017). Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2942–2950.

Kim, J., Rohrbach, A., Darrell, T., Canny, J., and Akata, Z. (2018). Textual Explanations for Self-Driving Vehicles. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, pages 563–578, Cham. Springer International Publishing.

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., Kästner, L., Schmidt, E., Sesing, A., and 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. and Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777.

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267:1–38.

Mohseni, S., Zarei, N., and 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., and Sharma, A. (2021). Towards Unifying Feature Attribution and Counterfactual Explanations: Different Means to the Same End. *AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 652–663.

Nyrup, R. and 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.

Preece, A., Harborne, D., Braines, D., Tomsett, R., and Chakraborty, S. (2018). Stakeholders in Explainable AI. *arXiv*, 1810.00184.

Raidl, E. and Rott, H. (2022). Towards a Logic for 'because'. *Unpublished Manuscript*.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 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.

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

Sokol, G. and Flach, P. (2020). Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67.

Tomsett, R., Braines, D., Harborne, D., Preece, A., and Chakraborty, S. (2018). Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. *arXiv*, 1806.07552.

Venkatasubramanian, S. and Alfano, M. (2020). The Philosophical Basis of Algorithmic Recourse. *Proceedings of 2020 ACM Conference on Fairness, Accountability, and Transparency*, pages 284–293.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2):841–887.

Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *International Conference on Machine Learning*, pages 2048–2057.

Zednik, C. (2019). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology*, 34(2):1–24.

Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C. (2019). Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philosophy & Technology*, 32(4):661–683.