Artificial Intelligence 297 (2021) 103498

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare ☆

Juan M. Durán

Faculty of Technology, Policy and Management, Delft University of Technology, The Netherlands

ARTICLE INFO

Article history:
Received 1 May 2020
Received in revised form 31 January 2021
Accepted 4 March 2021
Available online xxxx

Keywords:
Explainable AI
Scientific explanation
Medical AI
sXAI
Interpretable AI

ABSTRACT

Explanatory AI (XAI) is on the rise, gaining enormous traction with the computational community, policymakers, and philosophers alike. This article contributes to this debate by first distinguishing scientific XAI (sXAI) from other forms of XAI. It further advances the structure for *bona fide* sXAI, while remaining neutral regarding preferences for theories of explanations. Three core components are under study, namely, i) the structure for *bona fide* sXAI, consisting in elucidating the *explanans*, the *explanandum*, and the *explanatory relation* for sXAI: ii) the pragmatics of explanation, which includes a discussion of the role of multi-agents receiving an explanation and the context within which the explanation is given; and iii) a discussion on *Meaningful Human Explanation*, an umbrella concept for different metrics required for measuring the explanatory power of explanations and the involvement of human agents in sXAI. The kind of AI systems of interest in this article are those utilized in medicine and the healthcare system. The article also critically addresses current philosophical and computational approaches to XAI. Amongst the main objections, it argues that there has been a long-standing interpretation of classifications as explanation, when these should be kept separate.

1. Introduction

The implementation of AI systems in medicine is gaining rapid acceptance among health institutions, government agencies, and healthcare personnel. High-tech companies, such as IBM, are leading innovations in this field with systems such as Watson for Health. But while IBM presents this AI system as offering more objective medical decisions and more accurate diagnoses than actual oncologists [42,73], it has been pointed out that many of these claims have aggrandized the actual capacities of such systems [6,63,64,75]. Of particular concern is the case of IBM for Oncology, a state-of-the-art AI system capable of analysing large amounts of data and multiple variables with the purpose of rendering accurate diagnoses and treatments for cancer patients. It has been reported that this AI system has been unable to provide satisfactory justifications as to why a given recommendation was proposed. Proof of this can be found in the piloting of IBM's Watson for Oncology in Denmark and South Korea, where the outputs obtained were deemed to be largely deceptive on different accounts [79,31]. Particularly pressing was the fact that only a fraction of the outputs rendered by Watson for Oncology match – or closely match – the clinician's best diagnosis. As result, a large portion of the outputs would require further explanations if Watson

☆ This paper is part of the Special Issue on Explainable AI.
E-mail address: j.m.duran@tudelft.nl.

https://doi.org/10.1016/j.artint.2021.103498
0004-3702/© 2021 Published by Elsevier B.V.

for Oncology were to successfully diagnose a patient.[1] In view of these results, AI systems cannot fully participate in the practice of medicine without further offering basic epistemic functions, such as explanations of their outputs.

While IBM's Watson for Oncology was a notorious case, it is by no means an isolated occurrence. Many medical AI systems available today (e.g., Machine Learning, Deep Neural Networks, Support Vector Machines, etc.) present similar limitations in their capacity to explain and justify recommendations. Despite the disappointing results, medical AI is undoubtedly a promising technology that will constitute a major leap forward in healthcare practice and the medical sciences. For this reason, a chief challenge faced by today's medical AI is to be able to scientifically explain the outputs rendered by such systems (sXAI). Explainable artificial intelligence (XAI), as it is today approached and defined, focuses on designing automatic-decision systems that should be able to account for their recommendations and diagnosis to a human agent.[2] To this end, the current technology is delivering high-end classifications in the form of labelling, clustering, and pattern recognition that, presumably, facilitate the *interpretation* of the system and its outputs.[3] But classifications are not explanations, nor do they provide the same epistemic goods. What is still missing is an account of *bona fide* scientific explanations for medical AI, that is, a systematic study of the structure of explanations for medical AI. This structure must also be capable of providing the desired epistemic goods, such as scientific knowledge and understanding, conceptual coherence, and forms of influencing judgements, among others. Undoubtedly, medical AI is moving forward the practice and theory of medicine. But expectations must be managed if these systems are merely capable of offering highly accurate classifications of medical images and text over *bona fide* explanations.

This article offers a novel study on the logic of scientific explanations for medical AI. At its basis, the article inverts the current *bottom-up* models for a *top-down* approach. The standard strategy, pioneered by computer science and followed by philosophers (e.g., [80,47]), consists of structuring all forms of XAI attending to the current technology and available computational methodologies. But this strategy, I will argue, leads to confounding classifications with explanations. The *top-down* approach here advanced temporarily screens off the available technology and focuses instead on the conditions for a *bona fide* scientific explanation in medical AI. Admittedly, this approach might be, at first, technologically unfeasible. But at its core, it contains the necessary structure for channelling efforts into *bona fide* scientific explanations, the kind of explanation that medical AI needs.

The discussion proceeds as follows. Section 2 begins by briefly distinguishing scientific explanations from other forms of explanations. It then moves into arguing that current computational and philosophical approaches are classificatory rather than explanatory, and offers reasons as to why we must seek the latter. Section 3 focuses on fleshing out the structure for *bona fide* scientific explanation in medical AI. This section, then, is divided into two subsections. Subsection 3.1 deals with the standard units of analysis for scientific explanation. Here I discuss different possibilities for the unit that carries out an explanation (i.e., the *explanans*), the unit that will be explained (i.e., the *explanandum*), and the objective relation of dependency that links these two (i.e., the *explanatory relation*). Subsection 3.2 elaborates further on the role that human agents, medical practices, and non-epistemic beliefs have in explaining with medical AI. A *bona fide* sXAI cannot ignore relevant information such as the recipient of the explanation and the patient's values. Subsection 4 presents and discusses methodologies external to the AI system capable of measuring the *explanatory power* of, and human control over sXAI. Finally, section 5 briefly recaps the findings of this article and entertains the idea that the study carried out here sets the basis for further lines of research on sXAI.

2. The various forms of epistemic functions: classifications and explanations

Science advances our understanding of the world through diverse epistemic functions, such as explanations, predictions, classifications, and the like. The prime task of any theory of scientific explanation, the kind of epistemic function of interest to us, is to characterize the structure that accounts for and provides an understanding of the empirical world. Naturally, there is not one single type of scientific explanation. Well-understood explanatory pluralism informs us of the multiple and genuinely different explanatory practices across the sciences [43]. The notion of scientific explanation does suggest, however, two different contrasts. The first contrast is between explanations that are characteristic of science and those explanations that are not. This contrast revolves around the *demarcation* criteria of what constitutes a scientific enterprise, being the classic *locus* Popper's debate over *verificationism* and *falsationism* [53]. This first contrast has little bearing on the analysis of the structure of scientific explanation for AI that I intend to pursue here. It suffices to suggest that AI systems used in medicine, physics, and chemistry are somehow more scientific than the algorithm used for the user's preferences in an Amazon Wishlist. Indeed, medical AI makes use of objective desiderata accepted by scientific research, such as the use of scientific theories, the internal consistency and fitness to the relevant empirical data, the coherence with a well-established

[1] It is assumed the clinician's diagnosis is the benchmark against which the results are evaluated and accepted.

[2] Let us note that the literature waives between the AI system providing the explanation, and the AI system rendering enough information for humans to explain [52]. Here, I focus on the former interpretation.

[3] Guidotti and his colleagues offer a commonly found definition: "To *interpret* means to give or provide the meaning or to explain and present in understandable terms some concepts. Therefore, in data mining and machine learning, *interpretability* is defined as the ability to explain or to provide the meaning in understandable terms to a human [29]. These definitions implicitly assume that the concepts expressed in the understandable terms composing an explanation are self-contained and do not need further explanations. Essentially, an explanation is an 'interface' between humans and a decision-maker that is, at the same time, both an accurate proxy of the decision-maker and comprehensible to humans." ([30], 93:5).

body of scientific beliefs, and the fruitfulness for future research. Presumably, Amazon Wishlist clusters together pre-labelled items whose similitude is minimal. In this article, we shall keep these two forms of explanation separate.

The second contrast singles out explanations from other epistemic functions, such as predictions, classifications, and descriptions Hempel [33], and as such it holds a more prominent place in our discussion on sXAI. Of particular relevance, I submit, is the fact that much of what today is taken to be XAI are, in fact, classifications and predictions. But scientific explanations provide a particular type of valuable information, one that informs us about the world and grows our understanding of *why* a given output is the case, rather than organizing our knowledge and possibly forecasting new cases.[4]

The technical literature has devised an elaborated taxonomy for different types XAI [38]. At its core, two main approaches emerge.[5] The first one is *transparent box design*, which includes algorithms such as decision trees, Boolean rules sets, and generalized additive models. Transparent box design presupposes that these algorithms are completely surveyable by the pertinent agents (i.e., computer scientists, medical personal, and other stakeholders are able to follow the stepwise process of computing, inspect variables, etc.), and therefore explaining its outputs consists in showing the path-dependency of the algorithm that relates those outputs with the relevant function. This article is not interested in *transparent box design* for two reasons. First, because the algorithms suited for this approach do not have a strong presence in medical AI. Second, because there is already a philosophically-motivated logic for computer-based scientific explanation capable of accounting for these (see [22]).

The second approach is *post hoc interpretability*, which consists in providing an explanation to a black-box algorithm mediated by an *interpretable predictor* (i.e., a transparent algorithm). The function of the predictor is to make visible the internal workings of the black-box by showing the *path-dependency of the algorithm* that relates a given function (or set of functions) to its output. At its core, *post hoc interpretability* holds a similar strategy to *transparent box design*, with the main difference being that the latter is directly surveyable by humans whereas the former is by the interpretable predictor. In this respect, *post hoc interpretation* has a few interesting characteristics worth mentioning. First, they are model-agnostic, that is, there is no interest in ‘opening’ the black-box for explanatory, predictive, classificatory, or otherwise purposes. A second important feature is that *post hoc interpretation* is *transparency-conditional*. That is, explanations, predictions, and other epistemological achievements are bound to the mediating interpretative predictor, instead of the black-box itself. This means that, for *post hoc interpretation* to explain, there must exist a formal representation (e.g., isomorphism, partial-isomorphism, similarity, etc.) between the black-box model and the interpretable predictor. This representation is the one that warrants the success of the explanation.[6] Without it, there are no bases for claims that an explanation based on the predictor applies to the black-box algorithm. Unfortunately, the form of this representation is never spelled out.

This brief review brings forward the classificatory nature that I ascribed to XAI. Take *transparent box design* as a case in point. It is claimed that researchers explain by means of tracing back the outputs of the algorithm in the decision tree. Consider for instance the following recommendation: “the patient needs to ingest 50 mg of aspirin” Such an output is explained by locating the class to which the output belongs, say, “if the patient has headaches, then ingest 50 mg of aspirin.” This means that there is a traceable path-dependency between a function (or set of functions) and that output that corresponds to the internal workings of the algorithm. At its heart, path-dependency classifies outputs with respect to a given class. Likewise, *boolean decision rules* and *generalized linear rule models* are global *transparent box design* methods, since both are applicable for classification problems by implementing logical conjunctions (i.e., *and*-rules, *or*-rules, and systems of weights).

More complex examples include IBM AI Explainability 360 [38], which implements *contrastive explanation methods* as a local, *post hoc interpretability*. Interestingly, here we also have classifications of datasets based on diverse numerical properties, offering minimally sufficient, critically absent features of that dataset, and other desiderata. The surveyable and straightforward nature of these algorithms – or their exogenous mediators – make them transparent, and thus susceptible to establishing a sense of explanation and understanding. But, I submit, this is a false sense of explainability because classifications are not explanations.

Admittedly, the current state of technology is imposing serious restrictions on what can effectively be done with AI. And XAI is certainly not exempt from these technological limitations. The path-dependency that partisans of current XAI so eagerly defend is proof of these limitations: they can only account for *how* – or *that* – an algorithm reached a given output, but not *why*. So, if a medical AI offers the recommendation R = “the patient needs to take 50 mg of aspirin for the headache,” the system is essentially navigating the algorithmic path that leads to R. In fact, current XAI is only able to answer explanation-seeking *how*-questions, even when explanation-seeking *why*-questions are asked. For instance, “*why* does this medical AI suggest a dose of 50 mg of aspirin for a headache?” is answered by indicating *how* the system reaches that output – or, equivalently, *what* is its path-dependency. But at no point do current approaches provide answers to genuine explanation-seeking *why*-questions. To answer a *why*-question would consist in showing the causal-mechanical or

[4] Of course, classifications and predictions offer genuine forms of understanding, such as showing how new phenomena fit into well-established categories [78]. Although I will not pursue the question about the kind of understanding rendered by sXAI, there is generalized agreement that it is of a specific kind that largely comes from explaining the world [49,74].

[5] I am following Guidotti’s interpretation. *Post hoc interpretability* is also known as *post hoc explanation*, and *transparent box design* is also known as *direct interpretability*, among other names [38], [87]. For a formalization of *post hoc interpretability* and *transparent box design*, see ([30], 93:11 ff).

[6] Páez has brought attention to a similar issue in the context of understanding. To this author, “[t]he cognitive achievement reached by the use of these devices [i.e., *post hoc interpretations*] seems to differ in great measure from the understanding provided by an interpretative model” (2019).

inferential dependence between prostaglandin hormones and swelling, and the effect of aspirin in stopping the production of prostaglandin. That is, a supported or confirmed scientific explanation needs to be supplemented with objective relations of dependence in the system, such as the causal linkage between prostaglandin hormones and stopping swelling (e.g., through showing the acting biochemical components). Further considerations, such as the dangers of administrating aspirin to pregnant and breastfeeding women, must also be included in the explanation.

In this context, several questions slip into the cracks: is current XAI offering explanation-seeking *why*-questions or merely *how*-questions? Are classifications forms of explanation (e.g., are they providing the same epistemic goods)? Is surveying the path-dependency of an algorithm enough for claims about scientific explanations? What is the rationale for entrenching *bona fide* explanations? In what follows I offer a *top-down* analysis of the units involved in a *bona fide* explanation. Although I do not expect it to be technologically feasible *tout de suite* nor address every issue in connection with scientific explanations, I do expect it to shed light on the philosophical and technical debate on XAI. To accomplish this end, we need to begin with the recent philosophical attempt offered by Wachter and colleagues [80]. Owing to the limitations of space, my discussion will be brief and focus on two chief issues.

2.1. Counterfactual explanations for machine learning

Philosophers are showing an increasing interest in debates concerning XAI. A recent study can be found in Wachter and colleagues [80], who propose a *counterfactual model of explanation* for machine learning. This model has two chief virtues: first, it avoids any commitment to 'opening the black-box' by describing a dependency relation on external facts that lead to a decision made by the AI system ([80], 845). Second, it is computationally implementable under the current technology. In this context, the attempt by Wachter and her colleagues is a welcome move towards the analysis of sXAI. To my mind, however, counterfactual explanations fall short in several respects. To ground this claim, I briefly discuss two objections. First, contrastive explanations are too narrow, leaving out of consideration successful explanations. Second, counterfactual explanations, as elaborated by these authors, are classificatory and therefore the objections laid out in the previous section also apply here.

To marshal the evidence, counterfactual explanations are at their core contrastive, meaning that they require the generation of synthetic data points whose distance is minimal with respect to an original data point (i.e., the database that will be used by the AI system) ([80], 856). This, in turn, means that all synthetic data points are created from, and dependent on a pre-given dataset – otherwise, the system would be unable to measure the minimal distance. It must be expected, therefore, that any lack of information in the original dataset results in absences in the synthetic data set as well. But there are cases where the absence of information is, precisely, what explains a given pathology. For instance, the lack of sunlight explains the deficiency in vitamin D for some patients, which in turn explains hormone imbalance – particularly low levels of, or ineffective use of steroidal hormones [57,59,81]. In this sense, counterfactual explanations are too narrow, reporting on explanations that can only be traceable back to the original data set. The main problem with counterfactual explanations is that they are not required to cohere with a general body of scientific knowledge. Instead, the explanation is confirmatory of the outputs against a backdrop of information, which, incidentally, is at the basis of synthetic results.[7]

My second objection is that counterfactual explanations remain fundamentally classificatory. To illustrate this point, consider the example in Wachter et al.: 'What would have to be different for this individual to have a risk score of 0.5?'. The answer to this question is, "If your 2-Hour serum insulin level was 154.3, you would have a score of 0.51" ([80], 859). How does Wachter account for this explanation? A possible reconstruction – and blunt simplification – takes the variable *'2-Hour_serum_insulin_level'* as the counterfactual variable and *'score'* as the output variable. Then, the counterfactual explanation implements a conditional structure that relates the counterfactuals variable with the output score.[8] For instance, in the conditional structure of the algorithm, the counterfactual variable '154.3' relates to the output variable *score* '0.51.' Thus, the algorithm prints out the following counterfactual explanation: "If your 2-Hour serum insulin level was 154.3, you would have a score of 0.51". But as shown, this response is coded as a simple {IF... THEN... ELSE} conditional structure. But such a conditional structure has classificatory purposes, as it navigates the conditions under which '$score == 0.51$' are true, flagging all the values of the variable *'2-Hour_serum_insulin_level'* as counterfactuals for that world. In this respect, counterfactual explanations do not render information as to *why* this individual has a risk score of 0.5, but rather what needs

[7] This is, of course, not to say that there is some form of circularity with counterfactual explanations for AI. But unlike standard models of counterfactual explanation, socio-technical systems such as medical AI raise questions about the nature of the *explanans* and the *explanandum* when the former renders the latter [21,22]).

[8] It is important to mention that the authors leave unexplained how the counterfactual inference and counterfactual dependence can be epistemically grounded and technologically implemented, particularly in the context of a model of explanation that eschews (or so it appears) primitive assumptions on the causal structure of the world and the causal representation in the model. Wachter and her colleagues make explicit that their "approach does not rely on knowledge of the causal structure of the world or suggest which context-dependent metric of distance between worlds is preferable to establish causality" ([80], 848). Instead, it is claimed that "it will be more informative to provide a diverse set of counterfactual explanations, corresponding to different choices of nearby possible worlds for which counterfactual hold or a preferred outcome is delivered [...]" ([80], 848). Thus understood, providing a set of counterfactual explanations begs the question of how to ground the counterfactual inferences at the basis of the explanations, for these depend on assumptions about the causal structure of the world. The authors do provide, however, a list of considerations for the individual metrics of counterfactual explanations, including "the capabilities of the individual concerned, sensitivity, mutability of the variables involved in a decision and ethical or legal requirements for disclosure" ([80], 848). But more needs to be said. Lacking a satisfactory answer to these issues might cast doubts on the suitability of counterfactual explanations for medical AI.

to be altered in the variables of the algorithm to obtain the desired score. In other words, the counterfactual explanation informs *how* a given output was obtained, not *why*. Whereas in the latter case, we are demanding objective relations of dependence, in the former, we expect to convey the computational mechanisms that lead to a given output. If these considerations are correct, then Wachter's version of counterfactual explanations perpetuate the image of XAI as classificatory, and classifications are not explanations.

```
IF 2-Hour_serum_insulin_level ==[9] $\xi$
    THEN print "If your 2-Hour serum insulin level was" $\xi$
            ", you would have score of 0.51"
    ELSE print "Your 2-Hour serum insulin level is not" $2-Hour_serum_insulin_level
            ", please see your doctor."
```

A simple generalization in pseudo-code for a counterfactual explanation

3. Dissecting sXAI

Philosophy of science teaches us that scientific explanation takes the form of a well-defined structure, consisting of at least three units of analysis. Those are, the unit carrying out the explanation (i.e., the *explanans*), the unit that will be explained (i.e., the *explanandum*), and objective relations of dependence that link these two (i.e., the *explanatory relation*) [33,66,41]. Philosophers have also debated whether explanations should be objective or pragmatic. Objectivists typically consider that a fixed set of criteria is satisfied by all explanations. These criteria are universal in several respects: they apply to all scientific explanations, they do not incorporate specific empirical assumptions or presuppositions that might be made by scientists from different fields, and they are independent of the interests of particular audiences, among other criteria [83]. Pragmatists, on the other hand, eschew the objectivist's criteria and consider the psychology of the explainers and their audience (informed by the agent's beliefs, idiosyncrasy, use of language, desires, etc.) as playing a relevant role in the explanation itself. It follows that, to the pragmatist, explanatory traits such as explanatory asymmetries, the explanatory value of laws, etc. (see [66,40]) have their source in psychological facts pertaining to the explainer and the audience, rather than being independent of them (e.g., being ascribed to nature). Achinstein famously argued that "contextual explanations" are the kind of explanation that, given the provision of a specific body of information to some audience, will produce a sense of intelligibility conditional to the background knowledge and interests of that audience [1]. This means that explaining the basic principles of anatomy to trained physicians will diminish the epistemic appreciation of an explanation. Conversely, explaining complex models of the spread of COVID-19 to virologists will increase the understanding of the behaviour and dynamics of this particular virus.

The contrast between objectivist and pragmatist accounts of explanation can be used to shed light on the current debate on scientific explanation in medical AI, where players across the board seem too promptly to endorse some form of pragmatism. Take, for instance, IBM's AI Explainability 360 Open Source Toolkit, which differentiates algorithms and their applicability based on the 'audience' and the user of the algorithm:

> Global directly interpretable models are important for personas that need to understand the entire decision making process and ensure its safety, reliability, or compliance. Such personas include regulators and data scientists responsible for the deployment of systems. Global post hoc explanations are useful for decision maker personas that are being supported by the machine learning model [e.g., physicians, judges, and loan officers] (...) Local models are the most useful for affected user personas such as patients, defendants, and applicants who need to understand the decision on a single sample (theirs) [38].

Thus understood, IBM AI Explainability 360 permits the interests of the audience to enter into the structure of the explanation and play a relevant role in it, embracing the kind of pragmatism described earlier. I submit that this form of pragmatism is of the wrong kind for sXAI, leading high-end AI systems to misconstruct scientific explanations. To see my position, take the claim that different audiences have different explanatory needs. This means that the information required by a physician is, the claim goes, of a different kind from the information required by other personnel (e.g., nurses). Now, this bit of the claim seems undisputable. It is correct to say that, in a number of cases, nurses do not require the same explanatory information as physicians regarding, say, why a medical AI recommends chemotherapy over radiation therapy. My objection is to include the agents (e.g., physicians, nurses, and hospital managers) as irreducible facts to be referenced in the explanation, as it is required by pragmatism and adopted by IBM AI Explainability 360. To me, the structure of sXAI must remain unaltered irrespectively of who is giving or receiving the explanation (see my discussion in section 3.2).

Now, I admit that pragmatism has a strong case. The variety of purposes and complexity of medical AI systems, the differences in background knowledge of the practitioners, the diverse and changing practices of institutions, the patient's need, and a plethora of other causes offer no good reasons to believe that a single act of explaining will be equally successful

[9] The comparison operator could change according to different operators (i.e., <, >, etc.).

for all cases. Hence, we need a taxonomy of XAI conditional to the agent's psychology, the context where the explanation is taking place, and other pragmatic considerations. But this rationale unjustifiably conflates the analysis of the structure of explanations with the pragmatics of giving explanations. It is perfectly conceivable that the same explanation is shared by physicians, nurses, and hospital managers alike, even for cases where the information obtained from two separate acts of explanation varies. This is the viewpoint that I shall adopt here. To me, agents are merely the recipients of the explanation, not parts of it. Thus, their psychology is screened off, and the explanation conserves its own epistemically justified independent structure. In other words, the fact that different information must be delivered to different audiences has no bearings on the structure for a *bona fide* explanation.[10]

Here is when we must make a distinction between two senses of pragmatism: *pragmatism*$_1$ and *pragmatism*$_2$ [83]. At its core, *pragmatism*$_1$ takes considerations about the psychology of those involved in providing and receiving explanations as irreducible facts to be referenced in the explanation. *Pragmatism*$_1$ also takes into consideration the local context as irreducible for an explanation, such as the availability of resources, the specialization of the personnel, etc. [83]. The aim of any pragmatist theory of explanation is to combine the logic of explanation with considerations drawn from *pragmatism*$_1$.[11] By means of such a combination, the explanation offers an answer to a request for information relative to an audience and several contextual considerations. It follows that pragmatic theories of explanation might cut off objective relations of dependence (e.g., causal relations) and threaten the loss of valuable epistemic goods (e.g., objective scientific knowledge[12]), for they no longer occur independently of the audience nor the context. But to cite the psychology of an individual as part and parcel of an explanation is epistemologically irrelevant. This should be particularly true in medical AI, where we are interested in maintaining objective metrics of evaluation and measurement of the explanatory power of sXAI (see my discussion in section 4). In short, if an explanation is about offering answers relative to an audience, it becomes impossible to discriminate genuine explanations from those tinted by the values and the personal beliefs of the audience ([65], 21).

Pragmatism$_2$, on the other hand, takes pragmatic considerations such as the utility and usefulness of an explanation to be at the service of some goal connected to human interests (e.g., to advance scientific interests and understanding). Thus understood, *pragmatism*$_2$ avoids including psychological and contextual considerations as structural parts in the explanation, but considers them as serving pertinent epistemic purposes for an explanation. Thus understood, under the umbrella of *pragmatism*$_2$, the structure of explanations does not depend on irreducible psychological and contextual facts, but rather on objective relations of dependence that are embedded with pragmatic, non-epistemic information [83].

Taking stock of these two senses of pragmatism, I follow many philosophers of science in that there is a distinction worth making between the analysis of explanation and the pragmatics of explanation-giving ([65], 21). I now defend the claim that we can be explanatory objectivists for medical AI, while including elements from *pragmatism*$_2$. In this context, two core issues emerge. First, finding an objectivist motivated explanatory structure for medical AI that includes considerations drawn from *pragmatism*$_2$. Section 3.1 offers a *top-down* survey for such a structure. This section briefly presents and examines the three units of analysis for a *bona fide* sXAI. Recall from section 1 that a *bona fide* sXAI is understood as a logically well-structured scientific explanation for medical AI that provides the desired epistemic goals (e.g., scientific knowledge and understanding, fosters conceptual coherence, etc.). The section briefly discusses concrete concerns that emerge in the context of each unit of analysis. Second, attending to the pragmatic considerations mentioned earlier, section 3.2 discusses the role of the agent receiving the explanation and the context within which the explanation is given. It goes without arguing that I neither expect to be exhaustive nor to catalogue all possible agents and contexts.

3.1. A survey to the structure of bona fide sXAI

3.1.1. The explanandum: the idiosyncrasy of data and the factivity condition

A common position among partisans of XAI is to separate *local* from *global* interpretability. Global interpretability means that an explanation is about the inner logic of a model, one for which humans can "follow the entire reasoning leading to all the different possible outcomes" ([30], 93:6). Global interpretability is typically achieved by an interpretable and transparent model (i.e., an *interpretable global predictor*), one that is "able to mimic the behaviour of the black-box and it should also be understandable by humans" ([30], 93:12). Local interpretability, on the other hand, consists in explaining "a specific decision: only the single prediction/decision is interpretable" ([30], 93:6). In this sense, local interpretability consists of an interpretable local predictor that accounts for a given output of the system. In this article, I will only be interested in local interpretability.[13]

[10] Pragmatist theories of explanation face several theoretical difficulties that, I believe, sXAI cannot afford. See, for instance, [68].

[11] Considerations about the psychology of individuals as well as considerations about the context do not need to concur together in a pragmatic theory of explanation.

[12] There are several ways to interpret scientific objectivity. I follow Douglas in her interpretation that relates the trust I have in my own assessment with a strong sense of endorsing this trust for others. Thus, "when I call something objective, I am endorsing it for myself, and endorsing it for others. For example, when I call an observation "objective", I am saying that I trust the observation, and so should everyone else" ([19], 116. Italics in original.) This interpretation opposes the more subjective stand proposed by *pragmatism*$_1$.

[13] A further claim that this article cannot pursue at this point is that local and global interpretability do not differ in terms of their logic. Similar concerns emerge within the Hempelian Deductive-Nomological model of explanation, where it was possible to explain phenomena using laws, as well as the laws themselves ([33], footnote 33, 273).

Following local interpretability, the *explanandum* for XAI must be a diagnosis, a treatment, a dose to be administrated, a medical hypothesis, or any other linguistic or non-linguistic output rendered by the medical AI system. Naturally, these outputs come in different formats of data, including pictures, visualizations, matrices, vectors, propositions, etc. Now, since a *bona fide* sXAI requires entrenching objective relations of dependence, both the *explanans* and the *explanandum* need to share the same ontology (see my discussion on section 3.1.3). To illustrate this, consider two simple examples of explanation. First, say that we are interested in using causality to explain why the consumption of aspirin reduces headaches. For this, we need to show that acetylsalicylic acid acts upon the production of prostaglandin in such and such a way, and that such and such a causal relation is taking place. Contrast this type of explanations with mathematically deducing that $C_9H_8O_4$, the chemical representation of acetylsalicylic acid, stops the production of prostaglandin. For such an explanation to take place, we need to represent the production of prostaglandin by mathematical formulae. Both forms of explanations are *bona fide*, but only because the *explanans* and the *explanandum* share similar ontological characteristics. We would only be able to mathematically deduce causality in the case that the causal relations are mathematically represented (e.g., via causal models) [50].

Thus understood, the data format of the *explanandum* is, in principle, irrelevant for the structure of *bona fide* explanations, but not its ontology. This means that, regardless of the data format, a *bona fide* explanation needs to describe the *explanandum* in terms of a linguistic entity. In this context, several candidates emerge with force: propositions, arguments, and descriptions of facts and events are amongst the most common ([65], chapter V). Interestingly, choosing one ontology over another is not a trivial matter. Propositions, for instance, are required for deductive models of explanation, such as the Deductive-Nomological model, and therefore they require to reinterpret the data in the language of first-order logic [33]. Arguments, on the other hand, are more suitable for unificationist accounts, and as such, they can be interpreted as a rigorous description of a scientific claim that instantiates patterns of arguments [88], [40]. Whereas both are epistemic theories of explanation, propositions and arguments are logically different in terms of truth-preserving, reconstruction of logical and non-logical vocabulary, and the intentionality and extensionality of the terms involved (see [65], 160). It follows that different theories of explanations require different interpretations of the *explanandum*. Consider for instance counterfactual explanations as developed by Wachter and colleagues. These authors seem to support the idea that arguments are at the basis of the explanatory relation. Indeed, the reconstruction of a plausible *explanandum* is "What would have to be different in v_1, $v_2, \dots v_8$ for this individual to have a risk score of y?". Here $v_1, v_2, \dots v_8$ are eight different variables (e.g., number of pregnancies, age, and BMI) found in any standard medical textbook on women of Pima heritage at risk of diabetes ([80], 859). To me, adopting a description of the *explanandum* in the way suggested by the unificationist (and, if I am right about Wachter's interpretation, by counterfactual explanations too) is quite reasonable. In fact, it also squares well with many reconstructions of the outputs of AI systems: "Why does this mole of *size_x*, *shape_y*, *border_z*, and *concentration_of_melanin* is a melanoma?" [26,27] But of course, other ontologies might also apply (see my discussion in 3.1.3).

Under this interpretation of the *explanandum*, at least two core issues arise. First is the question of whether it is possible to explain any form of data, as opposed to explaining facts or events in the world [11,45]. The problem is that data is idiosyncratic in the sense that it is gathered, filtered, curated, and restructured before it is considered a piece of scientific knowledge – again, oppose this to observing real-world facts or events. It follows that an *explanandum* could be manipulated in different ways and for attending different purposes, losing in this way the objective stand that we deem so important to hold. For instance, an explanation involving a melanoma is no longer *bona fide* in the case that the image of the mole used for the explanation has been modified to fit that of melanoma.

A second concern that has some kinship with the idiosyncrasy of data is that the *explanandum* must be *true* in order to be explainable ([33], 247-248). This condition is known as the *factivity condition*, and it squares rather well with the intuition that we cannot explain what is false: we cannot explain *drapetomania*, a disease commonly diagnosed among American slaves in the 19th century, with the main symptom being the tendency to run away [13]. Another example is *spattergroit*, a disease transmitted by a fungus that covers the victim in purple pustules and renders them unable to speak. The reason why we cannot explain *drapetomania* and *spattergroit* is that neither exists. The former was invented to account for the very human desire for freedom, and the latter is a fictional disease from the Harry Potter saga. Similarly, to explain recommendations, treatments, doses, and so on rendered by medical AI, one needs to warrant their truth.[14]

An often-pursued solution to the factivity condition is to appeal to the *transparency* of the algorithm. Presumably, we could ground the truth of the output by showing the inner workings of the algorithm and how these workings relate to the output. But such a procedure would only show a link between the algorithm and its output, not between the output and some real-world data (that, in turn, has been credited as true). In this sense, transparency could certainly speak in favour of the robustness of the output of the algorithm (i.e., the fact that they have been rendered without significant computational artifacts), but not necessarily bridge the AI system to the real-world.

Whereas the idiosyncrasy of data is a topic with little gravitation in studies on XAI, the factivity condition has been brought up by the recent literature. Páez has recently sentenced that "machine learning is the kind of context in which one can say that, in principle, it is impossible to satisfy the factivity condition for understanding-why" [49]. I think Páez is right.

[14] I only use this term for consistency with the literature on explanation. However, to speak of 'truth' in medical AI is too ambitious, if not plainly wrong. I will not pursue this line of argumentation in this article. The interested reader is invited to browse through the literature on computer simulations [37,82,48,24].

But I also believe that the factivity condition could be tackled by ensuring the reliability of the AI system and the epistemic trust in their outputs (see the work on computational reliabilism [23] and [24]. But this is material for another occasion.

3.1.2. The explanans: path-dependency and close systems

Current treatments of XAI take explanation to be about showing the path-dependency of an algorithm to its outputs. Once this link is established, one has presumably explained these outputs. A typical problem that we face with medical AI is that they are black-box algorithms and thus cognitively opaque. The standard solution to this is to use an exogenous algorithm – called an *interpretable predictor* – that shows the path-dependency of the AI system. Thus understood, a melanoma is explained by showing that the medical AI classifies an image as having a size greater than 6 mm, an asymmetrical shape, and other characteristics that physiologically single out the melanoma [69].

Now, it has largely gone unnoticed that this form of explanation is not *bona fide*, for it depends on a *closed system* that produces the same outputs that it will later explain. Indeed, this arrangement for an explanation is analogous to taking the Ptolemaic model of planetary motion, calculating the orbits for several planets, and then explaining the apparent retrograde motion of these planets using deferents, epicycles, and the equant, all purposefully designed into the model. Have astronomers explained the motion of planets? Unlikely. "Explanation" is a success term that indicates that we have gained (possibly new) knowledge and understanding of the real-world. As presented, the astronomer has inferred the motion of the planets rendered by the model from the deferents, epicycles, and the equant already designed into the model. It follows that this explanation is not informing about the real-world, but rather about the model.[15]

A similar problem can be spotted in AI systems. Consider the controversial case of Wu and Zhang on automated classification of criminality using face image recognition [85]. This AI system was trained to identify criminals solely based on the traits and attributes of faces. According to the authors, the faces of law-abiding citizens have a higher degree of resemblance with each other, as opposed to the faces of criminals, which presumably have a higher degree of dissimilarity in facial appearance. Now, this example makes evident the problem with using an explanatory proxy whose only attribute is to show the inner workings of the algorithm. Any explanation as to why a given face is of a criminal – or not – is possible within this system; one only needs to find the right path-dependence that leads to that explanation-seeking outputs. But again, these explanations neither inform nor account for facts and events happening in the real-world.

For all intents and purposes, *bona fide* sXAI needs to account for outputs independently of any path-dependency in the algorithm. Since explaining a particular *explanandum* depends upon access to a suitable, integrated, and tenable body of scientific knowledge that bears on that *explanandum*, a sensible first approach to the *explanans* must also include sources of information relevant for accounting for the outputs of medical AI. That is, descriptions pertaining to the design-decisions about the algorithm, descriptions of sources of medical and biological evidence and theories, descriptions of background knowledge, descriptions of actual (and potential) correlations and causal relations, and descriptions of pragmatic considerations such as the agent's values and local institutional practices (see my discussion in section 3.2) are amongst my first choices. Let us remember that these descriptions must adopt a specific ontology conditional to the *explanandum*. It is, of course, conceivable that this list needs to be expanded. But it is only by conveying all the available relevant information about the AI system, the target system, the body of scientific beliefs, and pragmatic considerations that an explanation increases its epistemic value.

3.1.3. The explanatory relation: ontic and epistemic

Identifying the right explanatory relation is arguably the most challenging aspect of XAI. In the context of scientific explanation, a *bona fide* explanation depends on identifying objective relations of dependence between the *explanans* and the *explanandum*. By this I mean those structural features that characterize the *explanandum* and which are identifiable by scientific means obtain in virtue of being present in, and in connection to the *explanans*.[16] For instance, in the field of medicine, biology, and other related sciences, researchers identify the continuous mechanisms acting from cause to effect that connect the *explanans* with the *explanandum* (e.g., in a mechanistic setup [70,44,58], or within a causal nexus [67]). Alternatively, an epistemic approach is possible, one that entrenches relations of formal inferences between the *explanans* and the *explanandum* [33,40].

If we were to ask partisans of XAI what they have to say about the explanatory relation, we would probably hear very little. It is nevertheless conceivable that the computation involved in tracking the path-dependency of a black-box algorithm stands for some form of inference (presumably, a derivation). It follows that explanation *qua* path-dependency is within the family of epistemic theories of explanations. Now, although an interpretation of computation *qua* derivation has some initial appeal, it needs to be discussed in the larger context of the nature of computational processes [28,15,16,54], the relation between computation and cognition [55], the relation between (mathematical) proof and computation [4], the semantics of programming languages [76], and various studies on human reasoning and inference [3].

Wachter and colleagues offer an interesting alternative. These authors take the explanatory relation to be some form of counterfactual inference, as follows from adopting the counterfactual model of explanation. Unfortunately, the authors leave

[15] Of course, an explanation about the model and its inner working could be of value, for instance, to developers. However, we still need to distinguish such explanations from those that make claims about the world.

[16] It is worth taking note of philosophical discussions on dependency and grounding (e.g., [17]). Particularly important is to entrench the kind of relation of dependence holding between the *explanans* and the *explanandum* in algorithms. Unfortunately, this analysis cannot be done here.

unexplained how counterfactual inference and counterfactual dependence can be epistemically grounded and technologically implemented, particularly in the context of a model of explanation that eschews (or so it appears) primitive assumptions on the causal structure of the world and the causal representation in the model. In other words, the authors need to provide further arguments that ground the counterfactual dependence between the *explanans* and the *explanandum* (as opposed to computing a seemingly counterfactual output), and which are convincing in terms of their effective design and implementation on the computer (i.e., without resorting to a causal interpretation). Above and beyond these issues, Wachter and colleagues seem to be on a promising track when it comes to conceptualizing the explanatory relation as some kind of inferential (possibly non-causal) dependence. Several other scholars have taken a similar epistemic perspective. For instance, Aliseda [7] has pioneered philosophical work on abductive reasoning for AI, and Durán [22] has defended deductivist-unificationist accounts of explanation for computer-based systems such as computer simulations.

On more ontic grounds, many efforts are directed towards causal models [50,51,8]. Bengio, the 2019 Turing awardee, has called for more work on causal models that will pave the way for genuine forms of explanation for all AI systems [10]. Causal models certainly represent a promising approach to the explanatory relation in *bona fide* sXAI, provided of course that concerns about the notion of causality, inferences of new causal relations from piles of unstructured data [46], and non-causal explanations [60] are addressed.

3.2. The pragmatics of explanation

3.2.1. Multi-human agency and the recipients of the explanation

Scientific explanation for medical AI operates in an extensive background of professional knowledge, interests, and institutional regulations. The very same explanation could have several different recipients, and act for diverse purposes: physicians, nurses, hospital managers, none of which will presumably require the same explanation. Consider, for instance, a medical AI that recommends targeted therapy drugs consisting of a cocktail of vemurafenib, dabrafenib, and trametinib as the best treatment for advanced melanoma. One possible explanation is that this particular drug cocktail has proved to be highly effective among untreated advanced melanoma. Another explanation stems from its high success in cases with the specific conditions of the patient (e.g., the cancer cells have a particular genetic mutation). Yet another explanation points to the fact that alternative therapies (e.g., chemotherapy, radiation therapy, biological therapy) might affect the overall well-being of the patient, following-up with undesirable side effects (e.g., fatigue, hair loss, nausea) [9,32,56].

Let us say, as a working assumption, that all of the above explanations are *bona fide* explanations, that is, they are well-structured explanations that offer the right epistemological goods. The question that we must ask now is: who is the recipient of this explanation? Evidently, the first and second explanations go to the clinician, whereas the third explanation could also involve nurses and other non-medical personnel. It is indeed conceivable that nurses and other hospital personnel find the first and second explanations unnecessarily informative. Nurses might only need explanations pertinent to the right dose to be administrated. And hospital managers, much like any other hospital personnel not involved in medical care and the ways in which different drugs perform under a myriad of circumstances, only require explanations that account for the success rate of using a specific drug over another. This information might prove to be central for the negotiation with pharmaceutical companies over the price of a new batch of supplies. Conversely, physicians might very likely take that the explanation offered to nurses and hospital managers is epistemically meager and carry very little justificatory weight on how to decide over the prognosis of a disease or whether to accept a given treatment.

In this context, it must be expected that sXAI has a varying success rate. Clinicians demand some valuable information that, as a working assumption, deviates from the nurses' and hospital managers'. The specialized technical literature is quite familiar with problems of multi-agency and the nuances that come with it. The most commonly adopted solution requires implementing different explanatory structures conditional to the agent receiving the explanation (see, for instance, [38]). That is to say, human agency is a property of the explanation, and as such it must be included in the explanatory structure (i.e., in the sense given by *pragmatism*$_1$). No single XAI can account for multi-agency. The taxonomy proposed by IBM AI Explainability 360 is a good case in point (recall the quotation on page 9).[17]

To my mind, however, we can deal with multi-human agency without relinquishing objectivity in sXAI. That is, we treat the needs and desires of the recipient of an explanation as pragmatic considerations exogenous to the explanation itself. In short, we treat multi-human agency in the sense given by *pragmatism*$_2$. In this way, the explanatory structure for a medical AI remains unaltered regardless of the role of the agent receiving the explanation. It is instead the kind of descriptions used in the *explanans* that changes according to the needs of the agent (e.g., theories, data, models, etc.). Thus, if we manage to design a *bona fide* explanatory structure for medical AI, one that combines all the right components, then the very same structure can be used by physicians, nurses, and hospital managers, regardless of their individual epistemic goals. It is the information used in the *explanans* that changes as a function of the agent, not the structure of the explanation itself.[18]

[17] Another example is Arya et al. [5] where the authors showcase the multiplicity of explanatory systems available in the current market.

[18] I am not advocating for a universal structure of *bona fide* sXAI that encompasses all possible cases and all possible medical AI systems. Rather, the claim is that categorizing XAI based on agents and, as I will argue later, diverse contexts, introduces at least two undesirable implications: first, that XAI is subject to *pragmatism*$_1$, and the possibility of failing to deliver the right epistemic goods. Second, that multiplying XAI is not pragmatically attractive: would a hospital have different XAI systems in place, one per recipient of the explanation? We should thrive for unity in sXAI, to the extent possible, naturally.

To illustrate this, consider BenevolentAI, a medical AI that assisted British pharmacologist Peter Richardson in showing how rheumatoid arthritis drugs might significantly lessen some of the most severe effects of COVID-19 (BenevolentAI, 2020). BenevolentAI combines structured and unstructured biomedical data sources, drug industry data, and automated retrieval of information from scientific research papers. These sources are curated and standardized via highly sophisticated data analysis and data fabric. The company claims that "this is fed into our proprietary knowledge graph which extracts and contextualises the relevant information and is made up of a vast number of machine curated relationships between diseases, genes, [and] drugs" (BenevolentAI, 2020). Now, suppose that we want to know why BenevolentAI suggests rheumatoid arthritis drugs instead of any other drug in its catalog. The answer is that BenevolentAI relates the data analysis and data fabric to diseases, genes, and drugs through "curated relationships" (e.g., causal relationships, derivations). Under conditions of *bona fide* explanation, different explanation-seeking *why*-questions are answered by supplying specific chunks of information to the *explanans*, conditional to the recipient of the explanation. Thus, the pharmacologist explains why rheumatoid arthritis drugs are effective against COVID-19. The nurse, in turn, explains why the drug must be administered at −80 °C. And the hospital manager negotiates a new batch of drugs based on reasons as to why a particular drug is more effective, cheaper, and safer for patients. According to BenevolentAI, all these explanations are possible without altering its explanatory structure.

3.2.2. Embedding non-epistemic beliefs into explanations

Medical AI systems handle a myriad of data that go well beyond the science of medicine and biology. Some stems from the education and the practice of diagnosticians, surgeons, and emergency technicians, some comes from the patient values and personal decisions about tradeoffs between temporary alleviation and more aggressive treatment. The case of IBM's Watson for Oncology is paradigmatic in this respect. It has been reported to be highly successful for cases treated in the Memorial Sloan Kettering Cancer Center in New York, whereas the rate of agreement between Watson and oncologists drops to about 33 percent in places like Denmark [75,63]. This shows that Watson for Oncology is highly cohesive with local practices and the mores of care, which in no way can be deemed universal.[19]

The question that emerges in this context is the following: should *bona fide* sXAI exclusively require epistemic beliefs (e.g., medical and biological data, theories, etc. that account for the *explanandum*), or should non-epistemic beliefs, such as communal interests and personal values also be accounted for (e.g., via a description of local medical practices, the representation of patient's values, etc.)? To illustrate the complexity of this issue, consider the case of an oncologist interested in understanding why a certain mole was diagnosed as melanoma. In this scenario, an explanation is possible in the usual objectivist way: the *explanans* includes well-established models of melanoma, relevant data about the patient's physical condition, and past information about successful treatment of that kind of melanoma, among other epistemic beliefs. But such an explanation might be inadequate in a number of cases. The same oncologist might require explanations that need to include the myriad of treatments accepted and practiced in other institutions, as well as explanations that account for the patient's values and personal preferences. The concept of *bona fide* sXAI seems to exceed the idea of a structure that uniquely accounts for and embodies medical and biological theories and data. We must also account for the merits that normative beliefs bring into an explanation.

Under the new context, two familiar forms of sXAI emerge: one advanced by *pragmatism*$_1$, which includes the audience's psychology, intentions, background education, idiosyncrasies, and individual desires as irreducible facts to be referenced in the explanation. As before, such an approach carries undesirable implications, such as requiring a full-fledged taxonomy for XAI conditional to the institutional practices and patient's values. The alternative is to treat these non-epistemic beliefs in the sense offered by *pragmatism*$_2$, where the structure of explanations remains unchanged and the local practices and personal values offer some service connected to the agent's interests. For instance, an explanation for a specific audience (e.g., parents of a child) needs to be stripped of any technicalities, and an explanation for fighting individual pernicious prejudices must provide convincing information that a homeopathic treatment will not heal any type of cancer. Neither of these explanations requires a tailored-made structure, and both include institutional practices and personal values.

Unfortunately, the proposed solution to maintain objectivity faces a serious challenge. A particularly difficult case to frame is the patient's values. In some cases, they play a legitimate role in the explanation and therefore they can neither be dissociated from the explanatory structure nor disregarded as mere idiosyncrasies. Consider, for instance, a medical AI that detects some form of cancer in a patient and recommends chemotherapy as treatment. Say that the chemotherapy drugs have decreased the patient's ability to produce new blood cells. The medical AI system, therefore, detects blood cell counts dropping and recommends a blood transfusion as standard protocol for recovering the red cells and platelets count. Say, for the sake of argument, that the patient refuses the blood transfusion based on her personal values (e.g., religious, cultural, moral). What is, then, the epistemic status of the explanation? Should the medical AI have considered the patient's values as an indicator that chemotherapy would not have led to an acceptable explanation, or is the explanation itself *bona fide* insofar as it uses biological and medical theory and data, and it successfully establishes explanatory relations for that type of cancer? To put the same issue in more familiar terms: should the sXAI have incorporated the patient's values as part of the *explanans*, or should these considerations serve some pragmatic goal exogenous to the explanatory structure? This is not an easy question to answer, for we need to consider at least its epistemological and normative sides.

[19] Of course, data have also played their role in the system's failure to agree with oncologists, but data is always to blame.

My objectivist approach takes that personal values and other non-epistemic beliefs do not hold a specific weight in the explanation. This means that there are no reasons to think that an explanation as to why chemotherapy is the best option for a given type of cancer and for a given type of patient should be different because of the individual and personal values of the patients. The patient's values, at any rate, serve different pragmatic goals, such as motivating the physician to seek a different treatment (i.e., those that do not require a blood transfusion). But again, at no point are the epistemic merits of the explanation affected by the patient's values. This is particularly true if our interpretation of health is attached to statistical measurements and standards of normal biological functioning of the body [12]. Under this interpretation, a person is healthy or unhealthy based on objective biological measurements. But the concept of health could also be based on social constructions about what the relevant individuals or groups consider to be healthy [61]. Thus, a group that considers blood transfusion as harmful will treat any of its members who have received a transfusion as unhealthy [62]. For such cases, an sXAI that does not include this particular personal value in its structure might not be considered *bona fide*. If this were the case, then *pragmatism*$_1$ seems to be a more sensitive approach than my objectivist account. I could, of course, try to save my account by demanding that all non-epistemic beliefs are operationalized and implemented as part of the medical AI. But this solution does not come cheaply, as several known problems come in connection with embedding values in AI systems [77]. Unfortunately, I cannot pursue this issue any further.

4. Assessing the merits of bona fide sXAI: Meaningful Human Explanation

Thus far, I have portrayed *bona fide* explanations as a necessary condition for sXAI. But while *bona fide* explanations seek to entrench the right explanatory structure for medical AI, it remains silent on how strong and convincing a given explanation could be. To this end, we need to be capable of measuring the strength, convincing value, and influential utility that a particular *explanans* has over the *explanandum*. This is of central importance for sXAI, as physicians, nurses, and managers have limited resources for justifying their confidence in a given explanation. In what follows, I briefly explore *Meaningful Human Explanation* (MHE), an umbrella concept that intends to capture how humans could assess the epistemic merits of sXAI.

To my mind, MHE is achieved when at least three components are in place. First, we need a clear description of what constitutes the *explanatory power*[20] or *explanatory goodness* of sXAI. The work of Schupbach and Sprenger [72] and Schupbach [71] are instrumental to this end. These authors have advanced a Bayesian analysis of explanatory power that identifies objective and formal conditions of adequacy for a given theory of explanation. Interestingly, the authors claim that their analysis is uncommitted to any particular theory of (objective) explanation, a feature that squares well with my approach to sXAI. Equally important is the work of Ylikoski and Kuorikoski [86], who distinguish between five different dimensions of the goodness of an explanation, namely, non-sensitivity, precision, factual accuracy, degree of integration, and cognitive salience ([86], 208). To these authors, the goodness of an explanation is assessed in a given dimension by answering a range of counterfactual questions (i.e., what-if-things-had-been-different questions), along with showing their theoretical and pragmatic importance for the explanation.

Although both approaches offer very promising methods to measure the explanatory power of scientific explanations, further considerations need to be taken into account if we were to use them for measuring the explanatory power of sXAI. For instance, factual accuracy in Ylikoski and Kuorikoski's account should not be considered a decisive epistemic goal for sXAI. As suggested earlier, *truth* and data are seen by philosophers as two conflicting (possibly irreconcilable) concepts. Equally pressing is to operationalize Ylikoski and Kuorikoski's dimensions, or Schupbach and Sprenger's probabilistic function for its computational implementation. It is indeed a non-trivial matter of how concepts are computable.[21]

The second component of MHE aims to aggregate human control to the process of explaining. This is of course not to suggest a return to *pragmatism*$_1$. Rather, the claim is that humans are still a fundamental constituent in assessing whether explanations uphold epistemic and normative standards. The importance of human control over automated systems has gained attention in ethical studies of AI, but of course, it is not alien to social studies of science [19]. A chief concern here is that the design and use of AI systems are not value-free, as they involve social, ethical, and political considerations, not all of which necessarily align well with our values, mores, and beliefs. Consider for instance a Deep Convolutional Neural Network capable of classifying different skin lesions and identifying malignant cases [26]. The system has been showcased by its authors as a highly reliable medical AI, capable of accurately detecting, classifying, and ultimately demarcating melanoma from other forms of skin lesions. Despite the clear benefits that an automatic classification of moles could bring to the advancement of dermatological and oncological practice, as well as fulfilling the promise of universal access to healthcare, several voices have been raised against these systems and their methods. We find warnings that real-world detection and diagnosis involve a different set of normative assumptions from those designed and coded in the medical AI. For instance, humans adopt a precautionary principle which indicates that, in case of doubt, "err on the side of caution," even if this results in over-diagnosing malignancy [14]. Medical AI does not necessarily come with such safeguards built into their

[20] Schupbach and Sprenger analyze the sense of explanatory power as the "hypothesis's ability to decrease the degree to which we find the explanandum surprising (i.e., its ability to increase the degree to which we expect the explanandum)" ([72], 108).

[21] Studies on measuring the quality of an explanation in AI can also be found in [34–36]. These authors propose to implement the System Causability Scale (SCS) as such metric. The SCS consists of a ten-item questionnaire that measures "the quality of an explanation interface (human-AI interface) or an explanation process itself" ([36], 196). I thank an anonymous reviewer for pointing this literature out to me.

systems. Another concern is the multiple biases found in the training datasets that are adopted, "naturalized," and later used by the AI system: gender bias, racial bias, socio-economic bias, and the list continues [2,20,39]. Dermatologists are (or should be) immune to racial bias.

It is therefore paramount that humans do not lose sight of the explanations rendered by AI systems. This idea has been cogently articulated by de Sio and van den Hoven, who claim that "humans, not computers and their algorithms should ultimately remain in control of, and thus [be epistemically] responsible for, relevant decisions about [automated systems]" [18]. One way to achieve this end consists in agreeing on conditions of possibility under which an explanation is meaningful to humans. This can be done, in principle, by identifying concrete properties of an explanation that are relevant to humans' goals and susceptible to human evaluation. Potential candidates include ethical and social implications of an explanation (e.g., that an explanation enables discriminatory practices or morally justifies physicians (see [25]), and alternative interpretations of a concept (e.g., HIV serological assays in infants are difficult to interpret and thus open to different meanings [84]).

The third role reserved for MHE stems from evaluating the suitability of sXAI in the larger context of scientific beliefs and practices. In principle, a *bona fide* and explanatory powerful explanation has little value in and by itself. It is when an explanation with such traits makes possible reliable assessments of the merits of a new piece of knowledge, enables the integration of such knowledge into a larger body of medical and scientific beliefs, facilitates the learning and understanding of medical categories, fosters conceptual coherence, and allows for decisions about the suitability of one explanation over alternative explanations, among various other scientific practices that indicate that sXAI has been fully integrated into the scientific enterprise. Thus understood, the epistemic merits of an explanation do not (solely) come from being *bona fide* and explanatorily powerful. They also come from the myriad of ways in which explanations with such traits accommodate medical and biological knowledge and fosters the growth of scientific understanding. This side of any process of explaining with medical AI exceeds its automation. Human agents, again, prove to be irreplaceable epistemic actors.

With MHE we round out the structure of sXAI: we ensure that its structure is *bona fide*, and therefore the right epistemological goods are suitable to be delivered; we also ensure that we measure the explanatory power of a given sXAI, a key trait for reliable explanations; finally, sXAI must conform with humans' goals and standards, evaluations, and practices. If the explanation does not uphold ethical and social standards, then what good is it to be *bona fide* and explanatory powerful? Medical principles such as non-maleficence and beneficence indicate that the output just explained is unsuitable for the patient [25]. If the physician cannot make sense of an explanation by integrating it with prior knowledge, how valuable is the explanation? MHE brings the control and evaluation of sXAI back to humans.

5. Final remarks

To call an explanation scientific is not to say that it draws and articulates categories from a large body of scientific data. It is rather to say that it conforms to a specific, well-defined structure capable of advancing our understanding of the world. Explanations that follow such a structure are *bona fide*. This article holds via different claims that current studies on XAI fail to be *bona fide* explanations, erring on the side of being classifications. The article then advances a study of *bona fide* scientific explanation for medical AI by addressing three core components: the structure of sXAI (subdivided into the three units of analysis for any scientific explanation), the role of human agents and non-epistemic beliefs in sXAI, and finally how human agents can meaningfully assess the merits of an explanation. This article, then, proposes a shift from standard XAI to sXAI, accompanied by substantial changes in the way explanation in medical AI is constructed and interpreted. For this shift to effectively happen, it is paramount to look at what philosophy of science, epistemology, cognitive science, and other related disciplines have to say. Then – and only then – can we consider the computational implementation of sXAI.

Although this article takes pride in paving the way towards *bona fide* scientific explanation in medical AI, it also acknowledges that much more needs to be said. Each section only scratches the surface of much deeper issues that draw largely from philosophical studies of technology. But fully characterizing sXAI is an on-going endeavor that needs to be approached carefully in tandem with all the relevant parties.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Thanks go to Jack Casey and Giorgia Pozzi for reading this article in detail. All three reviewers deserve equal recognition for their close reading and helpful comments. One in particular needs to be singled out for keeping this author modest. Thank you. I would also like to thank the editors of the special issue for their encouragement and patience. Finally, I would like to thank the section Values, Technology and Innovation, Faculty of Technology, Policy and Management at the Delft University of Technology for their unrivalled support. This paper was partially done during my fellowship at the Netherlands Institute for Advanced Study in the Humanities and Social Sciences (NIAS-KNAW) under the project "Accountability in Medical Autonomous Expert Systems: ethical and epistemological challenges for explainable AI". This work was also supported by the European Commission through the H2020-INFRAIA-2018-2020 / H2020-INFRAIA-2019-1 European project "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (Grant Agreement 871042). The funders had no role in developing the research and writing the manuscript.

References

[1] P. Achinstein, The Nature of Explanation, Oxford University Press, 1983.

[2] A.S. Adamson, A. Smith, Machine learning and health care disparities in dermatology, JAMA Dermatol. 154 (11) (2018) 1247–1248.

[3] J. Adler, L. Rips, Reasoning. Studies of Human Inference and Its Foundations, Cambridge University Press, 2008.

[4] K. Arkoudas, S. Bringsjord, Computer, justification, and mathematical knowledge, Minds Mach. 17 (2007) 185–202.

[5] V. Arya, R.K.E. Bellamy, P. Chen, A. Dhurandhar, M. Hind, S.C. Hoffman, S. Houde, Q. Vera Liao, R. Luss, A. Mojsilović, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K.R. Varshney, D. Wei, Y. Zhang, One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques, arXiv:1909.03012, 2019.

[6] M. Aggarwal, M. Madhukar, IBM's Watson analytics for health care, in: C.M. Bhatt, S.K. Peddoju (Eds.), Cloud Computing Systems and Applications in Healthcare, IGI Global, 2017, pp. 117–134.

[7] A. Aliseda, Abductive Reasoning. Logical Investigations into Discovery and Explanation, Synthese Library, 2006.

[8] A.A. Altman, Causal models, summer 2019 edition, in: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, 2018, https://plato.stanford.edu/archives/sum2019/entries/causal-models/.

[9] A. Baldi, P. Pasquali, E. Spugnini (Eds.), Skin Cancer, Springer, 2014.

[10] Y. Bengio, An AI pioneer wants his algorithms to understand the 'why', Wired (10 August 2019). Accessed on 1 April 2020.

[11] J. Bogen, J. Woodward, Saving the phenomena, Philos. Rev. 97 (3) (1988) 303.

[12] C. Boorse, Health as a theoretical concept, Philos. Sci. 44 (4) (1977) 542–573.

[13] S. Cartwright, Report on the diseases and physical peculiarities of the negro race, in: A. Caplan, J. McCartney, D. Sisti (Eds.), Health, Disease, and Illness, Georgetown University Press, 1851 [2004], pp. 28–39, Reprinted (2004).

[14] R. Callen, J. Denny, M. Pitt, L. Gompels, T. Edwards, K. Tsaneva-Atasanova, Artificial intelligence, bias, and clinical safety, BMJ Quality & Safety 28 (2018) 231–237.

[15] T. Colburn, J.H. Fetzer, R.L. Rankin (Eds.), Program Verification: Fundamental Issues in Computer Science, Springer Science, 1993.

[16] T.R. Colburn, Philosophy and Computer Science, M. E. Sharpe Inc., 2000.

[17] F. Correia, B. Schnieder, Metaphysical Grounding. Understanding the Structure of Reality, Cambridge University Press, 2012.

[18] F.S. de Sio, J. van den Hoven, Meaningful human control over autonomous systems: a philosophical account, Front. Robot. AI 5 (2018).

[19] H. Douglas, Science, Policy, and the Value-Free Ideal, University of Pittsburgh Press, 2009.

[20] V. Dick, C. Sinz, M. Mittlböck, H. Kittler, P. Tschandl, Accuracy of computer-aided diagnosis of melanoma: a meta-analysis, JAMA Dermatol. 155 (11) (2019) 1291–1299.

[21] J.M. Durán, Explaining Simulated Phenomena: a Defense of the Epistemic Power of Computer Simulations, PhD thesis, Universität Stuttgart, 2014.

[22] J.M. Durán, Varying the explanatory span: scientific explanation for computer simulations, Int. Stud. Philos. Sci. 31 (1) (2017) 27–45.

[23] J.M. Durán, Computer Simulations in Science and Engineering. Concepts - Practices - Perspectives, Springer, 2018.

[24] J.M. Durán, N. Formanek, Grounds for trust: essential epistemic opacity and computational reliabilism, Minds Mach. 28 (4) (2018) 645–666.

[25] J.M. Durán, K. Jongsma, Who is afraid of black-box algorithms? On the epistemological and ethical basis of trust in medical AI, J. Med. Ethics (2021), forthcoming.

[26] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, S. Thrun, Dermatologist-level classification of skin cancer, Nature 542 (2017) 115–118.

[27] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, J. Dean, A guide to deep learning in healthcare, Nat. Med. 25 (2019) 24–29.

[28] J.H. Fetzer, Program verification: the very idea, Commun. ACM 37 (9) (1988) 1048–1063.

[29] Doshi-Velez Finale, Been Kim, Towards a rigorous science of interpretable machine learning, arXiv:1702.08608v2, 2017.

[30] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, F. Giannotti, A survey of methods for explaining black box models, arXiv:1802.01933, 2018.

[31] J.G. Hamilton, M.G. Garzon, J.S. Westerman, E. Shuk, J.L. Hay, C. Walters, E. Elkin, C. Bertelsen, J. Cho, A. Gucalp, A.D. Seidman, M.G. Zauderer, A.S. Epstein, M.G. Kris, A tool, not a crutch: patient perspectives about IBM Watson for oncology trained by memorial Sloan Kettering, J. Oncol. Pract. 15 (4) (2019) e277–e288.

[32] A. Hanlon, A Practical Guide to Skin Cancer, Springer, 2018.

[33] C.G. Hempel, Aspects of Scientific Explanation, Free Press, New York, 1965.

[34] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop?, Brain Inform. 3 (2) (2016) 119–131.

[35] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Mueller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9 (4) (2019).

[36] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: the system causability scale (SCS). Comparing human and machine explanations. KI, Künstl. Intell. 34 (2) (2020) 193–198.

[37] P. Humphreys, Extending Ourselves: Computational Science, Empiricism, and Scientific Method, Oxford University Press, 2004.

[38] IBM, IBM explainability 360, Technical report, IBM, 2019.

[39] T.B. Jutzi, E.L. Krieghoff-Henning, T. Holland-Letz, J.S. Utikal, A. Hauschild, D. Schadendorf, W. Sondermann, S. Fröhling, A. Hekler, M. Schmitt, R.C. Maron, T.J. Brinker, Artificial intelligence in skin cancer diagnostics: the patients' perspective, Front. Med. 7 (2020) 233.

[40] P. Kitcher, Explanatory unification and the causal structure of the world, in: P. Kitcher, W.C. Salmon (Eds.), Scientific Explanation, University of Minnesota Press, 1989, pp. 410–505.

[41] P. Kitcher, W. Salmon, Scientific Explanation, University of Minnesota Press, 1989.

[42] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, P. Vinck, Fair, transparent, and accountable algorithmic decision-making processes, Philos. & Technol. 31 (4) (2018) 611–627.

[43] T. Lombrozo, Causal–explanatory pluralism: how intentions, functions, and mechanisms influence causal ascriptions, Cogn. Psychol. 61 (4) (2010) 303–332.

[44] P. Machamer, L. Darden, C.F. Craver, Thinking about mechanisms, Philos. Sci. 67 (1) (2000) 1–25.

[45] J.W. McAllister, What do patterns in empirical data tell us about the structure of the world?, Synthese 182 (1) (2009) 73–87.

[46] V. McKim, S. Turner, Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences, University of Notre Dame, 1997.

[47] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in AI, Atlanta, Georgia USA, FAT*'19, 2019.

[48] M. Morrison, Reconstructing Reality. Models, Mathematics, and Simulations, Oxford University Press, 2015.

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

[50] J. Pearl, Causality. Models, Reasoning, and Inference, Cambridge University Press, 2000.

[51] J. Pearl, M. Glymour, N.P. Jewell, Causal Inference in Statistics. A Primer, Wiley, 2016.

[52] D. Pedreschi, F. Giannotti, R. Guidotti, A. Monreale, L. Pappalardo, S. Ruggieri, F. Turini, Open the black box data-driven explanation of black box decision systems, arXiv:1806.09936, 2018.

[53] K. Popper, Conjectures and Refutations, Routledge & Kegan Paul, 2002.

[54] G. Primiero, On the Foundations of Computing, Oxford University Press, 2019.

[55] G. Piccinini, A. Scarantino, Information processing, computation and cognition, J. Biol. Phys. 37 (2011) 1–38.

[56] E.R. Ranschaert, S. Morozov, P.R. Algra, Artificial Intelligence in Medical Imaging. Opportunities, Applications and Risks, Springer, 2019.

[57] J. Reiss, Third time's a charm: Wittgensteinian pluralisms and causation, in: P. McKay Illari, F. Russo, J. Williamson (Eds.), Causality in the Sciences, Oxford University Press, 2012.

[58] J. Reiss, Causality and causal inference in medicine, in: M. Solomon, J.R. Simon, H. Kincaid (Eds.), The Routledge Companion to Philosophy of Medicine, Routledge, 2016, pp. 58–70.

[59] J. Reiss, R.A. Ankeny, Philosophy of medicine, summer 2016 edition, in: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, 2016.

[60] A. Reutlinger, J. Saatsi (Eds.), Explanation Beyond Causation. Philosophical Perspectives on Non-causal Explanations, Oxford University Press, 2018.

[61] K.A. Richman, Ethics and the Metaphysics of Medicine, The MIT Press, 2004.

[62] K.A. Richman, A.E. Budson, Health of organisms and health of persons: an embedded instrumentalist approach, Theor. Med. Bioethics 21 (4) (2000) 339–352.

[63] C. Ross, I. Swetlitz, IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close, Statnews (2017), https://www.statnews.com/2017/09/05/watson-ibm-cancer/. (Accessed 25 April 2020).

[64] C. Ross, I. Swetlitz, IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show, Statnews (2018), https://www.statnews.com/wp-content/uploads/2018/09/IBMs-Watson-recommended-unsafe-and-incorrect-cancer-treatments-STAT.pdf. (Accessed 25 April 2020).

[65] D.H. Ruben, Explaining Explanation, Routledge, 1992.

[66] W.C. Salmon, Scientific Explanation and the Causal Structure of the World, Princeton University Press, 1984.

[67] W.C. Salmon, Causality and Explanation, Oxford University Press, 1998.

[68] W.C. Salmon, P. Kitcher, Van Fraassen on explanation, J. Philos. 84 (6) (1987) 315–330.

[69] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models, arXiv:1708.08296, 2017.

[70] K.F. Schaffner, Discovery and Explanation in Biology and Medicine, University of Chicago Press, 1993.

[71] J.N. Schupbach, Robustness analysis as explanatory reasoning, Br. J. Philos. Sci. 69 (1) (2018) 275–300.

[72] J.N. Schupbach, J. Sprenger, The logic of explanatory power, Philos. Sci. 78 (2011) 105–127.

[73] S. Somashekhar, Validation study to assess performance of IBM cognitive computing system Watson for oncology with Manipal multidisciplinary tumour board for 1000 consecutive cases: an Indian experience, Ann. Oncol. 27 (9) (2016).

[74] M. Strevens, No understanding without explanation, Stud. Hist. Philos. Sci., Part A 44 (2013) 510–515.

[75] I. Swetlitz, Watson goes to Asia: hospitals use supercomputer for cancer treatment, Statnews (2016), https://www.statnews.com/2016/08/19/ibm-watson-cancer-asia/.

[76] R. Turner, Understanding programming languages, Minds Mach. 17 (2007) 203–216.

[77] I. Van de Poel, Embedding values in artificial intelligence (AI) systems, Minds Mach. 30 (2020) 385–409.

[78] P. Verreault-Julien, How could models possibly provide how-possibly explanations?, Stud. Hist. Philos. Sci., Part A 73 (2019) 22–33.

[79] C. Vulsteke, M. Ortega Arevalo, C. Mouton, K. Stam, R. Goethals, F. Ameye, C. Populaire, M. Peeters, P. Verdonck, Artificial intelligence for the oncologist: hype, hubris, or reality?, Belg. J. Med. Oncol. 12 (7) (2017) 330–333.

[80] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the GDPR, Harvard J. Law Technol. 31 (2) (2018) 841–887.

[81] M. Weber, Causes without mechanisms: experimental regularities, physical laws, and neuroscientific explanation, Philos. Sci. 75 (2008) 995–1007.

[82] E. Winsberg, Models of success versus the success of models: reliability without truth, Synthese 152 (2006) 1–19.

[83] J. Woodward, Scientific explanation, winter 2019 edition, in: Edward N. Zalta (Ed.), The Stanford Encyclopedia of Philosophy, 2019.

[84] World Health Organization, Laboratory methods for diagnosis of HIV infection in infants and children, in: WHO Recommendations on the Diagnosis of HIV Infection in Infants and Children, 2010.

[85] X. Wu, X. Zhang, Automated inference on criminality using face images, arXiv:1611.04135v1, 2016.

[86] P. Ylikoski, J. Kuorikoski, Dissecting explanatory power, Philos. Stud. 148 (2) (2010) 201–219.

[87] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature 1 (May 2019) 206–215.

[88] P. Kitcher, Explanatory unification, Philos. Sci. 48 (4) (1981) 507–531.