Why and How Should We Explain AI?

Buijsman, Stefan

DOI
10.1007/978-3-031-24349-3_11

Publication date
2023

Document Version
Final published version

Published in
Human-Centered Artificial Intelligence - Advanced Lectures

Citation (APA)
Buijsman, S. (2023). Why and How Should We Explain AI? In M. Chetouani, V. Dignum, P. Lukowicz, & C. Sierra (Eds.), *Human-Centered Artificial Intelligence - Advanced Lectures* (pp. 196-215). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 13500 LNAI). Springer. https://doi.org/10.1007/978-3-031-24349-3_11

Why and How Should We Explain AI?

Stefan Buijsman(✉)

Department of Values, Technology and Innovation, Delft University of Technology,
Jaffalaan 5, Delft, The Netherlands
s.n.r.buijsman@tudelft.nl

Abstract. Why should we explain opaque algorithms? Here four papers are discussed that argue that, in fact, we don't have to. Explainability, according to them, isn't needed for trust in algorithms, nor is it needed for other goals we might have. I give a critical overview of these arguments, showing that there is still room to think that explainability is required for responsible AI. With that in mind, the second part of the paper looks at how we might achieve this end goal. I proceed not from technical tools in explainability, but rather highlight accounts of explanation in philosophy that might inform what those technical tools should ultimately deliver. While there is disagreement here on what constitutes an explanation, the three accounts surveyed offer a good overview of the current theoretical landscape in philosophy and of what information might constitute an explanation. As such, they can hopefully inspire improvements to the technical explainability tools.

Keywords: Explainability · Trust · AI ethics

1 Introduction

Artificial intelligence, and in particular machine learning methods, is fast gaining ground. Algorithms trained on large datasets and comprising numerous hidden layers, with up to a trillion parameters, are becoming common. Such models are difficult to explain to lay users with little understanding of the basis of machine learning, but they are also hard to interpret for those who designed and programmed them. The calculations that are carried out by the algorithm are not assigned an easily understandable meaning, aside from there being far too many of these calculations to actually follow. The outputs of algorithms are, as a result, hard to predict and to explain. Why did the algorithm output that there is a cat on this picture? We don't really know, certainly not without additional help in the form of explainability tools.

Reducing this vast range of calculations to an explanation that is accessible and helpful is a huge challenge, and one that is laid out in more technical detail in chapter [8]. Instead, I focus on the question: is it a problem that machine learning models are hard to interpret? This links closer to the tutorials on Ethics in AI, such as the chapters [2,15,20]. For there is a philosophical question at the very

M. Chetouani et al. (Eds.): ACAI 2021, LNAI 13500, pp. 196–215, 2023.
https://doi.org/10.1007/978-3-031-24349-3_11

start of any attempt to make machine learning models more explainable: *why do we need explainability?* And, furthermore, if we decide that explainability is important, what is it exactly that the tools we aim to develop should provide? What information is needed to constitute an explanation? When is a machine learning model interpretable?

These are big questions, and require far more discussion to be fully answered than I can provide here. So, to narrow the scope I primarily look at the recent increase in arguments against the need for explainable machine learning methods. I discuss the philosophical angles taken here, using four recent publications on the topic. I leave aside most technical work on explainability methods, for the interested readers see chapter [8] or the literature overviews in [1,4,10]. Instead I focus on philosophical accounts of explanations in the second part, presenting three different accounts of what explanations are and thus what information we should aim to provide in our explainability methods. As such, this chapter should give a more conceptual overview of reasons we have for wanting to explain machine learning/AI algorithms and what that might entail.

1.1 Learning Objectives

At the end of this chapter readers will have:

- Awareness of the recent arguments against the need for explanations of machine learning algorithms
- Ability to discuss the merits of these arguments
- Knowledge of the main accounts of explanation in the philosophical literature

2 Why Should We Explain Algorithms?

With the growing use of machine learning algorithms, and the benefits they may deliver, there has also been an increase in discussions about the conditions in which we can use these algorithms. Do we require algorithms to be explainable or is it acceptable to use algorithms even when we can't say why they produce the outputs they do? I discuss four different (sets of) arguments that aim to establish that explainability is not necessary to trust, or acceptably use, an algorithm. They help set up a discussion of the aims of explainable AI (XAI) tools and cover the philosophical literature up to 2021. While this leaves out some discussions from computer science [18] in the same period, it gives a good overview of the arguments that circulate against a requirement of explainability. Do we need explanations for trust in AI? [7] What about other goals we might have for such tools? Are they such that all, or most, machine learning algorithms require XAI support? The four papers have been chosen to reflect these questions: from arguments against needing explainability for trust (2.1 – 2.3) to arguments considering the wider range of goals explainability might serve (2.4).

2.1 Robbins: The Impacts of AI and a Catch-22

Robbins [17] would certainly answer this last question with a resounding 'no'. He presents two arguments for this conclusion, but starts with the idea that any requirement on explainability wouldn't attach to the algorithm as such, but rather to the decisions that are made with/by the algorithm. For example, we don't require people to be generally explainable; it's fine if some actions (e.g. a sudden urge to dance by a child) are left unexplained. Likewise, I don't need to explain why I crave Chinese food for dinner, though I do need to offer an explanation were I to hit someone. It's the impact, presumably, of these different decisions on other people that makes all the difference. If the impacts are, or can be, high, then we require explanations. If they are low and nobody would be affected by my decisions, then I can make them however I like. So Robbins [17] argues, and so I'll assume here to trace out the argument.

For it is what happens in the different cases where it gets interesting. The low-risk situation, where an algorithm's outputs aren't really capable of leading to harms, are supposed to be ones where we don't need explanations. As with people, where decisions of little import to others can be made without needing a good reason for them, so algorithms too shouldn't have to be explainable. Robbins [17] uses the example of AlphaGo, the choices of which cannot be readily explained. The lack of any risks associated with AlphaGo's outputs means that nobody considered this problematic, though perhaps we would find it valuable additional information. The point is rather that the lack of potential harms means that there is no *requirement* of explainability. If the algorithm fulfills its function well enough, it should be acceptable to use it.

So far the argument is uncontroversial. Proponents of a requirement of explainability typically point to high-stakes cases such as algorithms making medical diagnoses, as it is in these situations that we would expect explanations from people. Here Robbins [17] discusses a 'Catch-22' argument to show that requiring machine learning algorithms to be explainable would make them redundant. The idea is that "in order for the explanation given by explainable AI to be useful we must have a human capable of knowing which considerations are acceptable and which are not. If we already know which considerations are acceptable, then there is no reason to use ML in the first place. We could simply hard-code the considerations into an algorithm - giving us an automated decision using pre-approved, transparent, reasoning." [17, p.510] Explanations can only be used to calibrate trust if we know what good reasons for a decision are, and Robbins claims that if we have that knowledge then we could just as well program a transparent, rules-based, algorithm.

It's this last claim that I consider problematic, as we often do seem to be capable of evaluating decisions even if we can't write down the exact steps needed for making or even evaluating the decision. The sheer amount of work by cognitive scientists to figure out how e.g. our visual processing functions, is a good example of this. We are very good at identifying objects and telling whether someone has correctly done so and on the correct basis, but formulating general rules that capture all this (e.g. defining when an object is a cat) turns out to be very difficult. We're bad at giving necessary and sufficient conditions, and at providing dictionary definitions even though we know perfectly well what our words mean and which objects are cats. The same holds for less conceptual tasks such as driving, where we can tell whether someone is a good driver, and whether decisions were made for appropriate reasons, but have a hard time giving a full functional description of how we drive. The very reason we started with machine learning was this observation that it is extremely difficult to formulate exact rules that capture our cognitive abilities. So the claim that the ability to evaluate explanations is sufficient to formulate transparent GOFAI (Good Old-Fashioned AI) algorithms seems blatantly false. As a result, the argument for the redundancy of explainable machine learning algorithms fails. For even if we can evaluate explanations (tell whether the algorithm looks at the right things to decide whether something is a cat), that doesn't imply that we could specify a rules-based algorithm instead. Explainability can still be a reasonable requirement on machine learning algorithms, at least in situations where the potential impact is serious enough.

2.2 London: A Lack of Explanations Elsewhere

If the reason that we get out of the Catch-22 in [17] is that we often can't provide the detailed descriptions needed for a transparent algorithm, then isn't that lack of transparency on our side a reason to question a requirement of explainability? Aren't we requiring something of algorithms that people wouldn't be able to supply, when we claim that they should be interpretable? This is the direction of critique in London [14], who claims that "The opacity, independence from an explicit domain model, and lack of causal insight associated with some powerful machine learning approaches are not radically different from routine aspects of medical decision-making." [14, p.17] The examples that London uses to make this point are important for the evaluation of the argument. For he points out that:

> [M]odern clinicians prescribed aspirin as an analgesic for nearly a century without understanding the mechanism through which it works. Lithium has been used as a mood stabilizer for half a century, yet why it works remains uncertain. Large parts of medical practice frequently reflect a mixture of empirical findings and inherited clinical culture. In these cases, even efficacious recommendations of experts can be atheoretic in this sense: they reflect experience of benefit without enough knowledge of the underlying causal system to explain how the benefits are brought about. [14, p.17]

In other words, we might not always be able to explain why a certain treatment is effective, even if it is the responsible decision to prescribe that treatment. London reinforces that point with an example of treatments that were chosen based on theoretical reasoning, which was later disproved (and showed that the chosen treatments in fact harmed patients). "In medicine, the overreliance on theories that explain why something might be the case has sometimes made it more difficult to validate the empirical claims derived from such theories, with disastrous effects." [14, p.18] It is the efficacy of the treatments, verified in randomized clinical trials (RCT) that offers the best basis for deciding which treatment to offer. And so, London argues by analogy, we should decide which algorithms to trust/use based on their efficacy and not based on whether we have a thorough understanding of them.

The question is whether this comparison to our (in)ability to explain why treatments work is a good one. For despite the fact that we may not know the underlying causal mechanisms, doctors are clearly able to explain why they choose to prescribe a certain treatment: it is, for example, the experience that it has worked for patients with similar symptoms, ideally backed up by RCTs. Similarly, they will have reasons available (that can be evaluated by other medical professionals) to support their diagnosis. It is therefore important to consider the role that machine learning algorithms are supposed to play. They are not a type of treatment about which decisions are made, but a tool to aid or replace decision making, much closer to the diagnostic tools such as MRI machines that are used and whose functioning we can explain in detail. In the case of decision making we can, and frequently do, receive explanations.

Now, this doesn't quite settle the matter, as we can also ask whether we trust our doctor and machinery (and consider a diagnosis and treatment decision a good one) based on an ability to explain a decision, or based on a track record of previous good decisions. Perhaps here there is a similar case to be made that what we value is their reliability, rather than their ability to give reasons for decisions. The point to make regarding London's argument, however, is that the lack of understanding of causal processes that explain the efficacy of treatments is largely besides the point. These treatments don't constitute decisions or diagnoses. There is no rational deliberation here, nor is there a question of what a decision was based on. Instead, the cases he highlights show that making decisions for the wrong reasons (theoretical ones rather than evaluated efficacy) can be damaging, and so that having these reasons can be an important part of figuring out how good a decision is. Perhaps, without meaning to, this gives us precisely a reason to prefer or require explainable AI so that there too (automated) decisions can be shown to be based on the relevant features, considered in an appropriate manner.

2.3 Durán and Jongsma: Reliabilism and Explainability

Still, I consider the question whether there is a requirement of explainability on machine learning algorithms to be far from settled. London's examples may not give as strong an argument against such a requirement as he hoped, there are more perspectives from which to proceed. A more principled one is that of reliabilism, a philosophical position in epistemology that (in a very simple, process reliabilism formulation) analyses our justification to believe a proposition to be true in terms of the reliability of the process that led to this belief [9]. If the process is reliable (enough), then the belief is justified. If it isn't reliable, then there is no justification. How much of this we are aware of, or the internal reasons we have, is of little matter: the only thing that counts is the (externally determined) reliability of the process.

It is with this account in mind, though with a somewhat more involved notion of reliability than standard in the epistemological literature, that Durán and Jongsma [6] argue against a requirement of explainability on machine learning algorithms. While their main point is that (computational) reliabilism can offer an account of when to trust machine learning algorithms, they also give an argument against explainability. “transparency is a methodology that does not offer sufficient reasons to believe that we can reliably trust black box algorithms. At best, transparency contributes to building trust in the algorithms and their outcomes, but it would be a mistake to consider it as a solution to overcome opacity altogether.” [6, p.2] The underlying issue, as they see it, is one of infinite regress: our explainability tools will never succeed in making machine learning algorithms transparent, because of the need for transparency in the tools themselves. In their words:

> To see this, consider P again, the interpretable predictor that shows the inner workings of A, the black box algorithm. The partisan of transparency, S, claims that P consists of a sequence of procedures of which a given p_i entails A (or some of its outputs). But what reasons does S have to believe this? All that S holds is a very appealing visual output produced by P, like heatmaps or decision trees, and the - still unjustified - belief that such an output represents the inner workings of A. For all S knows, P is as opaque as A (eg, it can misleadingly create clusters which are biased, it can ignore relevant variables and functions that compromise the integrity of the results, etc). It follows that all we can say is that P induces on S the belief that S knows the output of A (ie, the idea that A is transparent), but at no point P is offering genuine reasons to believe that S has interpreted A. For this to happen, for S to be justified in believing that A is transparent, P must be sanctioned as transparent too. The problem has now been shifted to showing that P is transparent. [6, p.2]

There are two related concerns here that one might have about explainability tools. First, the tool itself might not be transparent, so e.g. the way in which a particular heatmap is output may not be explained. Why does the tool provide is with this explanation of the output rather than with another explanation? In other words, tool P may be a black box algorithm that in itself requires additional tools for its workings to be explained. Second, it might not be the tool itself that is opaque, but rather the fidelity of its explanations. We may know perfectly well how P generates explanations, but be unsure whether the explanations by P are indeed correct descriptions of why algorithm A reached that output. Linear approximations, for example, might be transparent in the first sense, but evaluating whether the explanations offered by them are a good representation of the reasons the machine learning model made its decisions is far more difficult. I think that Durán and Jongsma [6] aim to point to the second worry, about the fidelity of P, rather than the first, also since explainability tools often do not involve machine learning and so are generally not opaque in the first sense.

Would a regress occur here? Clearly it would if the first case was true: if all explainability tools are themselves black boxes, we wouldn't manage to get rid of black boxes by using XAI tools. That isn't the case, so they need a regress on determining the fidelity of these explainability tools. Is there some way to determine the fidelity of an explainability tool where we are sure that the answers are correct, or at least don't need yet another method to determine their correctness?

This will depend on the explainability methods we look at. Indeed, decision trees and linear approximations have this question of their fidelity to the model. And it is difficult to interpret heatmaps correctly, as there is a tendency to conceptualize the highlighted areas (e.g. a shovel) in terms of the concepts we typically use, whereas the algorithm might use the colour, shape, or some more complex combination of factors to make its decisions. While heatmaps highlight which areas are important for the algorithm's output, they do not explain *why* these areas are important. Yet the fidelity of these explainability tools is not directly in question: heatmaps, and feature importance methods in general, correctly represent part of the algorithm's behaviour. Similarly, counterfactual explanations correctly identify which minimal changes (under a chosen distance function, which is transparent – or can be transparent) lead to a change in output. Granted, these methods do not manage to make the algorithm transparent, but it is not their lack of fidelity that leads to this issue. I see, therefore, no reason to suspect a principled regress here on the question whether explainability tool P correctly describes the behaviour of algorithm A. Yes, we haven't found tools yet that manage to truly explain black-box algorithms (and probably this is why there are so many attempts to argue that algorithms can still be used even if they aren't explainable), but that alone is insufficient reason to think that it is impossible to achieve transparency.

Of course, Durán and Jongsma [6] offer a more constructive account of when people are justified to trust a machine learning algorithm, using reliabilism. It is not my intention to evaluate this account, on which explanations aren't needed for trustworthy algorithms, here (primarily for reasons of space). Rather, I aim to consider whether the arguments against a requirement of explainability hold up. Is it possible to hold that explainability is needed for trust? So far it seems that the answer is yes, as the arguments against such a requirement seem to fall short. It may still turn out that computational reliabilism, which they defend, is the correct account, but there is more room to argue for transparency than they suggest. In the case of an argument for explainability as a pre-requisite for trust, but also for an explainability requirement due to other reasons. It is this broader view that I explore in the next subsection, with the arguments of Krishnan [12] who goes through a range of possible ends to which explainability is a means.

2.4 Krishnan: Reasons to Explain

There are various reasons we might have to require, or strive for, explainability of machine learning algorithms. There is, as she terms it, the 'justification problem', closely related to questions about trust in algorithms. Explainability is often seen as a way to improve trust, and the above attacks on explainability have tended to focus on this aspect of the question (though Robbins actually focuses on control in his text). Her response here, as that of Durán and Jongsma, is to point to reliabilist accounts in epistemology. There are well-respected theories on which justification isn't linked to interpretability or some kind of awareness of reasons, and only links to reliability. So, putting forward a strong requirement of explainability motivated by trust is neglecting such alternatives, and might be wrong if reliabilist epistemologies are correct.

Trust/justification in general might then be in place without transparency. But a closely related reason to strive for explainability, or perhaps even require it in high stakes contexts, is often overlooked in these discussions. Yes, the algorithm might be trustworthy in general, because it gets it right often enough. But how about our trust in individual outputs? We might be justified to believe these, but surely we'd prefer to have a way to tell whether the algorithm makes good decisions in individual cases. For example, algorithms will perform worse in outlier cases, even when they are generally accurate. Understanding why the algorithm presents a certain output may help to gain a more fine-grained justification, and avoid mistakes that would arise if the algorithm is given blanket trust.

Krishnan would probably respond to this that there are other methods to detect such situations, such as outlier detection methods, which do not require the algorithm to be explainable. At least, that is the argument structure she uses for the other goals that explainability might serve. I can only agree here, such tools are available, and we have a fairly good idea of which conditions impact the reliability of our algorithms. But here, too, we can't make distinctions between outliers where the algorithm makes decisions in a way that seems reasonable, and ones where the algorithm makes decisions in an outlandish manner. For which outliers did the model generalize correctly, and for which will it make gross mistakes? Outlier detection won't be able to tell us. More insight into the processing of the algorithm might, however, provided that we have enough domain knowledge to evaluate its functioning. So there might actually be epistemic reasons to prefer explainability which cannot be readily replaced by accuracy measures. I'll leave this question aside though – I suspect that it's in part an empirical matter whether people actually are better at spotting mistakes if they understand why the algorithm gives a certain output – to move on to the other reasons one might have to demand explainability.

The second such reason that Krishnan discusses is that of fairness, or anti-discrimination in her terms. She naturally agrees that we should tackle this problem, but argues that explainability might not be a necessary ingredient to do so effectively. Fairness measures, possibly supported by synthetic datasets, might do most of the work here. Further investigations into biases of the training data can also help diagnose the source of discrimination by algorithms. Identifying and solving this issue, she argues, is something we can do largely by examining outcomes and training data. The algorithms needn't be transparent for that. I think that's a fair conclusion to draw, though as some of the participants in the workshop pointed out it doesn't mean that explainability methods will not help. Nor is there a guarantee that fairness measures, synthetic datasets, and so on will in practice spot all problems. Perhaps we can do this more effectively if algorithms are transparent than if they are not. For practical reasons, then, we still might prefer explainable algorithms over black boxes for this reason. Time will tell whether it is needed or not, but as she says "Interpretability is more plausibly construed as one tool among many which may prove useful against discrimination." [12, p.497]

The third reason that Krishnan puts forward is that one might want explainability to help weigh the output of the algorithm against other sources of information, or to solve disagreements. Generally speaking, it is the question of "how to synthesize the outputs of different sources into an overall decision and corresponding level of confidence in that decision" [12, p.497]. This is fairly close to the point made above, about the question whether we can trust individual decisions and the role that explainability may have to play there. Krishnan, however, focuses specifically on disagreements and balancing the output of an algorithm with other sources of evidence. It's a related case, but one that focuses primarily on disagreements. Her take on such disagreements is that here, too, we can do without detailed knowledge of why the algorithm reaches a certain output:

> When human processes and ML processes track different indicators, they are independent sources whose coincidence should strengthen confidence in the accuracy of a shared conclusion, whereas tracking the same indicators (and especially using the same indicators in a different way) can strengthen confidence in the aptness of the way that human reasoners are processing a given data set, without providing fully independent evidence in favor of the accuracy of that conclusion. Both scrutiny of the content of training data sets and ways of testing classifiers to see what features they actually track are viable ways of extracting this information without scrutiny of the steps that the algorithm performs in arriving at categorizations. [12, p.498]

Having this knowledge of features will certainly help, though I worry that resolving disagreements will not be as straightforward as this. When the same indicators are used, but different conclusions are reached, we surely want to know which way of considering the indicators is better. When people are given the same case, but disagree, we ask them why they reach their respective conclusions, to evaluate which arguments make the most sense. Similarly in the case with the AI, we would want to know whether the disagreement implies that we should trust the algorithm less, or the human judgement. Is one of the two clearly mistaken, or are there good reasons to put forward for both sides, which leads to suspension of judgement in this case of disagreement? Having this information about indicators is useful, but it alone doesn't tell us which confidence levels need to be adjusted downwards in cases of disagreement. Other measures, such as whether the case is dissimilar to the training data of the algorithm, might help here but won't give as much clarity as when the reasons for the output can be evaluated.

Similar issues arise if the algorithm has considered the same indicators, but agrees with us. Should this raise our confidence in the aptness of people's reasoning? Surely this depends, for example on whether the algorithms output is biased (we wouldn't want to say that the person's biased reasoning is apt because it coincides with a similarly biased AI output) or based on spurious correlations. Granted, the question of bias was tackled earlier, but here too we see that whether we raise our confidence or not depends on the type of process that led to the AI output. Krishnan nicely points out that we might learn a lot about this proces without it being transparent, and without explainability tools. Can we really learn all that we need to handle disagreements and reconcile evidence from different sources though? If we look at disagreements with people, then they do seem to resort to answers to why-questions in order to settle disputes. We ask people for their reasons to hold something true or false. Merely resorting to which factors were considered does not seem a sufficient substitute for that, though we might get further than we originally thought should it prove too difficult to make black box algorithms interpretable.

Furthermore, there might be other reasons to require interpretability of (some) algorithms. Krishnan lists two at the end of her paper [12, §3.4]. First, using machine learning in scientific contexts, to generate causal explanations and make scientific discoveries. Second, interpretability might be needed for public trust in these systems, even if it isn't necessary to be justified in believing the outputs. Perhaps, she reasons, society simply won't accept opaque machine learning models in some contexts, partly due to experts who favour explainability requirements. I, and the attendants at the workshop, think we can readily add a number of other goals to which explainability might contribute. Contestability is one, and is part of the reason that counterfactual explanations were introduced [21]. Explanations might also help with accountability, control and model debugging, to name a few.

I should stress, however, that this does not mean that I have now given a positive argument to *require* explainability in certain contexts. All I have room to add here is that the arguments surveyed aim to show that a common practice among people – giving reasons for actions – can be replaced in the AI setting by input features and outcome measures. I'm skeptical that we'll be able to do everything with these substitutes that we can with reasons. Evaluations of the merits of individual decisions, for example, are harder to do even if we can get quite a lot out of knowing what features were used and how reliable the algorithm is. Likewise, figuring out what to do in disagreements between algorithms or between algorithms and people is harder without reasons. One may argue that what we can do without reasons is good enough, and that explainability methods would be a great addition but not a necessary one. An answer to such an argument should depend on more specific cases, with the stakes laid out and our tolerance for mistakes, alongside the capability to detect and handle such mistakes with and without explainability methods. In setting medical diagnoses, for example, we might point to the serious impacts of mistakes and thus the desire for accountability mechanisms and a way to improve after a mistake has been made. Explainability can be important as a result, although it is to be carefully weighed against potentially reduced accuracy. Credit scoring algorithms will present outcomes that are still very impactful, but less so than the life-and-death situations of some medical decisions. Consequently the balance may shift there. In addition, Krishnan and others show that explainability is not a silver bullet: factors such as reliability, bias detection and the interaction between humans and the system all matter for responsible use of AI. Explainability might help, but likely will not solve everything.

In fact, it might turn out that *current* explainability methods don't deliver understanding of AI outputs, and that we also don't want to accept the use of black box algorithms. The conclusion, then, would be that we should work on developing interpretable algorithms instead of black box algorithms. For an argument to that effect, I refer the reader to [18]. It's an interesting argument about what we should give most priority to in current development practices, but one I'll leave aside here. As in the workshop, I want to focus on explainability methods for black box algorithms, and the question of how we should explain algorithms, if we agree that there are good reasons to strive for such explanations. I turn to that in the next section.

3 What Constitutes an Explanation?

The question posed here, 'what constitutes an explanation?' is deliberately abstract. For it is not my goal to offer concrete guidelines on how existing explainability tools are to be improved. Instead, I want to offer some theoretical perspectives that may be unfamiliar to those working in the field of explainable AI, yet may prove fruitful. Philosophers, namely, have thought about the nature of explanation (primarily in the context of the natural sciences) for at least a century. They have tried to unravel the structure that explanations follow, and theorized what it is about those bits of information that provides us with insight. As we strive for similar insight into the workings of machine learning algorithms, it may be that these different views of what information makes up an explanation offer a goal on the horizon to work towards.

One place to start is that in philosophical discussions there is a fairly widespread agreements that whatever explanations are exactly, they are answers to *contrastive* why-questions. The idea is that when we ask for explanations, we typically do so by asking 'why?'. When scientists want an explanation for a phenomenon, they ask why it occurred. When we want an explanation for an action, we ask why that person did it. However, the idea is that these questions have a more specific focus than just the phenomenon or action; they seek an explanation for a specific aspect of it. As Peter Lipton puts it in his influential book on inference to the best explanation:

> We do not explain the eclipse tout court, but only why it lasted as long as it did, or why it was partial, or why it was not visible from a certain place.

> Which aspect we ask about depends on our interests, and reduces the number of causal factors we need consider for any particular phenomenon, since there will be many causes of the eclipse that are not, for example, causes of its duration. More recently, it has been argued that the interest relativity of explanation can be accounted for with a contrastive analysis of the phenomenon to be explained. What gets explained is not simply 'Why this?', but 'Why this *rather than* that?' (Garfinkel1981:28–41; vanFraassen 1980:126–9). A contrastive phenomenon consists of a fact and a foil, and the same fact may have several different foils. We may not explain why the leaves turn yellow in November simpliciter, but only for example why they turn yellow in November rather than in January, or why they turn yellow in November rather than turn blue. [13, p.33]

That in itself doesn't tell us what the answers to contrastive why-questions look like, but it does point to something that might be missing from current explainability methods. Those do not offer a mechanisms for interest relativity, nor is there an option to consider explanations as operating with both a fact to be explained and a foil for which this explanation should operate. As a final example from Lipton to drive home the relevance of such contrasts: "When I asked my, then, 3-year old son why he threw his food on the floor, he told me that he was full. This may explain why he threw it on the floor rather than eating it, but I wanted to know why he threw it rather than leaving it on his plate." [13, p.33]

Granting that such contrasts are relevant, and that explanations might be construed as answers to contrastive why-questions, the natural follow-up question is: what do such answers look like? Here opinions differ among philosophers, and as a result I discuss three different accounts of what explanations are. I start with the kind of causal account that Lipton hints at, and has become more prominent since under the label of interventionism (and sometimes also manipulationism). Second, I discuss the unificationist framework of scientific explanation, and finally the mechanist framework of such explanations. I stick to brief overviews of these, and won't be arguing for any of the three frameworks over and above the other. Instead, I hope that this section can act as inspiration to further explore these rich discussions of explanations.

3.1 Causal/Interventionist

One possible answer to contrastive why-questions is to offer relevant causes. Why do leaves turn yellow in November rather than in January? Well, we can say something about the causal process that makes them turn yellow, and how this depends on the temperature drops that we already see in November. Not only is that a kind of answer we give routinely, it also matches nicely with the idea that explanations are asymmetrical: we can e.g. explain the length of a shadow based on the length of the object that casts the shadow. It would strike us as strange, however, to explain the length of an object by talking about how long the shadow it casts is, and what the position of the sun is at that time.

Causes can help account for this asymmetry, as it is the object that causes the shadow to be there, and not the other way around.

And so philosophers, such as Woodward [22] have tried to fill in the details of how causes might act as explanations. He starts out defining causation (or rather, a way to determine whether something is a cause, as the definition relies on a previous understanding of cause) on the basis of interventions. The idea is, basically, that X is a cause of Y iff an intervention on X that changes the value from x_1 to x_2 leads to a corresponding change in the value of Y, from y_1 to y_2. The important part here is how we understand interventions, where I is an intervention-variable on X with respect to Y if and only if:

> I1. I causes X.
> I2. I acts as a switch for all the other variables that cause X. That is, certain values of I are such that when I attains those values, X ceases to depend on the values of other variables that cause X and instead depends only on the value taken by I.
> I3. Any directed path from I to Y goes through X. That is, I does not directly cause Y and is not a cause of any causes of Y that are distinct from X except, of course, for those causes of Y, if any, that are built into the I-X-Y connection itself; that is, except for (a) any causes of Y that are effects of X (i.e., variables that are causally between X and Y) and (b) any causes of Y that are between I and X and have no effect on Y independently of X.
> I4. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X. [22, p.98]

Using this understanding of causes, it is possible to define minimal conditions on when a cause acts as an explanation E for explanandum M. The basic idea behind this definition is that causal explanations should give us information on the value that variable Y actually takes (e.g. the actual output of the algorithm) and inform us how this value changes should the value of X change. Which changes we're interested in depends on the contrast we focus on in our why-question. That then leads to the following more formal definition:

> Suppose that M is an explanandum consisting in the statement that some variable Y takes the particular value y. Then an explanans E for M will consist of
> (a) a generalization G relating changes in the value(s) of a variable X (where X may itself be a vector or n-tuple of variables X_i) and changes in Y, and
> (b) a statement (of initial or boundary conditions) that the variable X takes the particular value x.
> A necessary and sufficient condition for E to be (minimally) explanatory with respect to M is that (i) E and M be true or approximately so; (ii) according to G, Y takes the value y under an intervention in which X takes the value x; (iii) there is some intervention that changes the value of X

> from x to x' where $x \neq x'$, with G correctly describing the value y' that Y would assume under this intervention, where $y' \neq y$. [22, p.203]

While this definition really is rather minimal – a single counterfactual case will do – it showcases the underlying idea of causal explanations. The role of pointing to causes is so that you can determine what would happen if the situation (e.g. the inputs) had been different. Causes are simply those things that affect the outcome/output, and therefore the relevant factors to present. The goal of explanations, on this interventionist account, is thus to be able to tell how changes in the input result in changes to the output. Note that this is thus rather different from the explainability tools that are known as 'counterfactual', as those focus on presenting a single case in which the outcome is different. They do not give us a generalization that covers the actual case and a range of counterfactual cases.

If there's one thing to take away from this account of explanations, then it is the focus on the question: what happens if things are different? Good explanations should answer a wide range of such questions, on this account, and doing so with higher accuracy is naturally better. Furthermore, the range shouldn't be restricted to cases where the outcome is the same (as currently often happens for rule-based local explanations in XAI), but should cover how the outcome/output changes too. In a way this is quite similar to the other two accounts, as the unificationist for example focuses on deriving as many outcomes as possible with the same explanation, though the focus on causes is specific to the interventionist. With that, I leave causes aside, and move to the next account where derivations are the central element of explanations.

3.2 Unificationist

The basic idea behind this second account of explanation, unificationism (defended e.g. by Kitcher in [11]) is that the main role of explanations is to unify a range of different phenomena/observations. For example, Newton's laws of motion gave us a far better understanding of the physical world not because they point to causes, but because they presented us with formulas that showed how a range of different phenomena (e.g. different forces) all behave in the same way. They give a unified way to derive motions on Earth and in the heavens (stars, planets), something which was until that point unavailable. Gravity was derivable as simply another force, of which falling motions on Earth were but one example. It is such an ideal of unification, where a theory manages to give a single set of rules that bring together a wide range of phenomena, that Kitcher holds as the paradigmatic example of succesful explanations.

Yet, what does it mean to unify different phenomena (in the AI case that would be different outputs)? To make this more precise, Kitcher appeals to the notion of argument patterns that can be filled in different ways. The idea is that we start with sentences that have variables in them, that can take different values. Such sentences are schematic:

> A schematic sentence is an expression obtained by replacing some, but not necessarily all, the nonlogical expressions occurring in a sentence with dummy letters. Thus, starting with the sentence “Organisms homozygous for the sickling allele develop sickle-cell anemia,” we can generate a number of schematic sentences: for example, “Organisms homozygous for A develop P” and “For all x, if x is O and A then xis P” [11, p.432]

To turn a schematic sentence back into a ‘normal’ sentence, you need to add filling instructions, which specify what values the variables in the schematic sentence take. For example, A can be specified as taking the value ‘allele’ in the above schematic sentence. Once there are filling instructions for all variables in the schematic sentence, it is no longer schematic and simply a sentence. This functioning of schematic sentences allows for the construction of more abstract argument patterns, which form the basis of the unificationist account. With a little more setup, the basic idea is that we can go through the same procedure for arguments:

> A schematic argument is a sequence of schematic sentences. A classification for a schematic argument is a set of statements describing the inferential characteristics of the schematic argument: it tells us which terms of the sequence are to be regarded as premises, which are inferred from which, what rules of inference are used, and so forth. Finally, a general argument pattern is a triple consisting of a schematic argument, a set of sets of filling instructions, one for each term of the schematic argument, and a classification for the schematic argument. [11, p.432]

If we can find a single argument pattern (like Newton’s laws of motion) that allows for a very broad set of derivations, then we’ve succeeded in the goal of unifying phenomena. In Kitcher’s words: “Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same pattern of derivation again and again, and in demonstrating this, it teaches us how to reduce the number of facts we have to accept as ultimate” [11, p.432]. Unificationists, then, care about the number of outcomes/outputs that can be correctly derived from the explanation (argument pattern) that is offered. There is no particular focus on cases where the outcome is different, as with the interventionist framework of explanation. Nor do the argument patterns have to involve causes. Of course, it might turn out that general argument patterns often appeal to causes, but the unificationist can be rather more flexible here in what can figure in an explanation.

As a result, the unificationist framework might be closer to the current explainability methods. Of course, local explanations using decision rules/trees don’t manage to unify very many outcomes, and aren’t always correct in their predictions of model behaviour. But the basic direction is one that a unificationist, looking for patterns from which different outcomes can be derived, might endorse. There are more ways to go, however, as one might look not just at unifying different outputs of the same algorithm, but strive for argument patterns that work across algorithms or for connecting the outputs of one algorithm with theories in that domain of application. At least in the case of healthcare, Durán has made an argument for precisely such a connection with scientific theories in the explanations of algorithms [5]. Drawing more of these connections may be a way to push explanations forward. Although it may also help to, instead of unifying with parts outside of the algorithm, look at the inner mechanism more. That, at least, is the idea behind the last account I discuss.

3.3 Mechanists

Finally, there is a third account that again looks at causes to some extent. It differs from interventionism, however, in that the real focus here is on models that capture the mechanisms of a system. Those models are what do the explaining, and it is the representation of mechanisms in virtue of which these models explain. These accounts have changed somewhat compared to when it was first presented by Salmon [19], as they now look at constitutive explanations rather than etiological ones:

> Etiological explanations reveal the antecedent events that cause the explanandum phenomenon. Constitutive explanations, in contrast, reveal the organized activities of and interactions among parts that underlie, or constitute, the explanandum phenomenon. More specifically, they describe features of the mechanism for a phenomenon, where the mechanism includes the set of all and only the entities, activities and organizational features relevant to that phenomenon. [3, p.297]

This is a view of explanation that seems to work quite naturally in, for example, biology where if we want to understand the human body we're interested in a model that captures the functioning of the different parts and their causal role in the whole. The difficulty in applying this conception to AI is that machine learning models are far too complex for a complete model of their workings (which we have, after all) to offer much insight. And it's not just machine learning where this is the case, as we can wonder for the human body too how detailed these mechanistic models should be. Do they have to include absolutely everything that is relevant? Are more details automatically better? The point of [3] is to argue that this isn't the case, and that only those factors that are causally relevant matter. This is reflected in why they think that mechanistic models explain:

> A constitutive mechanistic model has explanatory force for phenomenon P versus P' if and only if (a) at least some of its variables refer to internal details relevant to P versus P', and (b) the dependencies posited among the variables refer causal dependencies among those variables (and between them and the inputs and outputs definitive of the phenomenon) relevant to P versus P'. [3, p.401]

Note again the focus on a contrast between P, which we want to explain, and P' from which it needs to be distinguished. Such contrasts can limit which parts of the mechanism are relevant. Furthermore, not every detail matters: only those that are causally relevant do. The question is where this leaves us with machine learning models, as here every parameter might have some causal relevance for the output of the algorithm. Can we really restrict the model to a manageable set of factors, then, in the case of machine learning algorithms? Perhaps not if we look at explanations of the internal mechanism for the algorithm itself, but if we take a slightly broader view of the system at play there is room for applying the account to AI. The training set, hyperparameters, number of layers, etc. all influence the outcomes, namely. They are causally relevant, and there aren't so overwhelmingly many of them that a complete mechanistic model would fail to generate real insights. An option, then, to apply the mechanistic account to AI is to take the factors not on the level of individual parameters (of which there are millions to trillions), but on a slightly more abstract level. It still makes sense to speak of a mechanistic model, and all the different factors are causally relevant for the outcome. So, an explanation in this sense looks at that interplay of factors, and as such offers a somewhat different way of considering causation in the explanation of AI outputs.

3.4 Bringing Theory into Practice

The types of explanations found in the philosophical literature are far from what current XAI tools provide. So, what can developers of (X)AI tools do about this discrepancy? First of all, as can be seen explicitly in most accounts, explanations are ideally presented in a contrastive format. That is, they answer questions of the form 'why P rather than Q?' instead of plainly 'why P?'. As pointed out elsewhere [16] this is a common feature of how humans explain, yet is not found in current XAI tools. Incorporating contrasts, and using these to direct the explanation, can be a first step towards XAI that more squarely fits the philosophical definitions of explanation.

Second, while current XAI tools might not conform to the definitions of explanation, that does not entail that they can serve none of the goals of explainability. Feature importance maps and counterfactuals can hint that the model looks at features that we think are irrelevant, even if they generally do not enlighten us as to why the model looks at those features. Similarly, bias detection can be aided by current XAI tools if part of the features the model uses are protected. In practice, therefore, XAI tools can still be useful even if they do not provide explanations proper.

Finally, I hope that the definitions provided here can help in the development of new explainability tools that do get closer to explanations as conceptualized in philosophy. A focus on generalizations that include counterfactuals (a combination that is rare in current XAI methods) is suggested by both causal and mechanistic accounts of explanation. In addition, Tools or explanations that consider what happens if training data or hyperparameters are changed, that to my knowledge do not yet exist, are also suggested by these accounts and might help us understand how we can improve AI systems. Most of all, the challenge is to design tools that can highlight the patterns used by AI models, as the unificationist account also points out. That is easier said than done, given both the non-linearity of machine learning models and their complexity. Still, clear definitions of explanation have been lacking in the literature on XAI and here philosophy can be helpful.

4 Conclusion

I have given a brief overview here of the reasons put forward against a requirement of explainability, and of accounts that might help us fill in what this end goal of explainability might mean. I've argued that providing such explanations doesn't make opaque machine learning methods redundant, as we typically can evaluate reasons without being able to replicate our thinking in decision rules. It also doesn't have to be unfair to ask for explanations in this case, as the examples pushed by London aren't of decision making processes, but rather focus on the treatments about which we decide. That brought us to reliabilism, and the question of whether there isn't some kind of regress in asking for explainability. While I think that such a regress is unlikely, there is a question of whether explainability is truly a requirement for trust/justification. Perhaps reliabilism offers an account on which we can do without explainability, and perhaps we can do more with input/output measures than we thought – as Krishnan points out. Still, there seem to remain good reasons to develop explainability methods, even if we don't end up requiring algorithms to be transparent.

Because of those reasons I went on a brief tour through accounts of explanation in the philosophical literature. For if we aim to make machine learning explainable, it might help to have a better sense of what explanations would ideally look like. Views differ here, but there is a surprising amount of agreement about the contrastive nature of explanations. One that, surprisingly, we do not yet see in technical explainability tools. Proceeding from there we see a range of options: interventionists who focus on causal dependencies and the ability to say what the output would be if the input were different. Unificationists who look at argument patterns that are as general as possible, allowing us to derive a wide range of outcomes. And finally mechanistic accounts, who focus on models that describe the inner workings of a system with as many causally relevant details as possible. Hopefully these can inspire improvements to explainability methods.

5 Suggested Readings

- Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead [18]
- Miller, Explanation in artificial intelligence: Insights from the social sciences [16]

- Woodward, Making things happen: A theory of causal explanation [22]
- Kitcher, Explanatory unification and the causal structure of the world [11]
- Craver and Kaplan, Are more details better? On the norms of completeness for mechanistic explanations [3]

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018)
2. Boella, G., Mori, M.: An introduction to ethics and AI. In: Chetouani, M., et al. (eds.) ACAI 2021. LNCS, vol. 13500, pp. 245–260. Springer, Cham (2022)
3. Craver, C.F., Kaplan, D.M.: Are more details better? on the norms of completeness for mechanistic explanations. Br. J. Philos. Sci. **71**(1), 287–319 (2020)
4. Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (XAI): a survey. arXiv preprint. arXiv:2006.11371 (2020)
5. Durán, J.M.: Dissecting scientific explanation in AI (sXAI): a case for medicine and healthcare. Artif. Intell. **297**, 103498 (2021)
6. Durán, J.M., Jongsma, K.R.: Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI. J. Med. Ethics **47**(5), 329–335 (2021)
7. Ferrario, A., Loi, M.: How explainability contributes to trust in AI. SSRN 4020557 (2022)
8. Giannotti, F., Naretto, F., Bodria, F.: Explainable machine learning for trustworthy AI. In: Chetouani, M., et al. (eds.) ACAI 2021. LNCS, vol. 13500, pp. 175–195. Springer, Cham (2022)
9. Goldman, A.: What is justified belief? In: Pappas, G.S. (ed.) Justification and Knowledge, pp. 1–23. Springer, Dordrecht (1979)
10. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Computi. Surv. (CSUR) **51**(5), 1–42 (2018)
11. Kitcher, P.: Explanatory unification and the causal structure of the world (1989)
12. Krishnan, M.: Against interpretability: a critical examination of the interpretability problem in machine learning. Philos. Technol. **33**(3), 487–502 (2020). https://doi.org/10.1007/s13347-019-00372-9
13. Lipton, P.: Inference to the Best Explanation. Routledge, Milton Park (2003)
14. London, A.J.: Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent. Rep. **49**(1), 15–21 (2019)
15. Methnani, L., Brännström, M., Theodorou, A.: Operationalising AI ethics: conducting socio-technical assessment. In: Chetouani, M., et al. (eds.) ACAI 2021. LNCS, vol. 13500, pp. 304–321. Springer, Cham (2022)
16. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019)
17. Robbins, S.: A misdirected principle with a catch: explicability for AI. Mind. Mach. **29**(4), 495–514 (2019). https://doi.org/10.1007/s11023-019-09509-3
18. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019)
19. Salmon, W.C.: Scientific explanation: three basic conceptions. In: PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, pp. 293–305. no. 2. Philosophy of Science Association (1984)

20. Slavkovik, M.: Mythical ethical principles for AI and how to attain them. In: Chetouani, M., et al. (eds.) ACAI 2021. LNCS, vol. 13500, pp. 275–303. Springer, Cham (2022)
21. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv. J. Law Tech. **31**, 841 (2017)
22. Woodward, J.: Making Things Happen: A Theory of Causal Explanation. Oxford University Press, Oxford (2005)