Pragmatic Explanations for Algorithmic Decisions

a dissertation for the completion of:
MSc Philosophy of Science

London School of Economics and Political Science

2022

Candidate Number: 32236

Supervisor: Dr. Laurenz Hudetz

Word Count: 9,509

Abstract:

As algorithmic decisions continue to play increasingly consequential roles in our lives, it is reasonable that we are offered explanations in some cases. First, then, there must be an account of algorithmic explanation to evaluate if something *is* an explanation, and the quality evaluated thereafter. Most accounts of explainable artificial intelligence offer a somewhat narrow view– a specific explanation type or mechanism, which by itself doesn't seem to be a general account. I argue for a pragmatic account of explanation, based on van Fraassen's writing. Crucially, this makes an explanation context-dependent and gives wide scope to what can be considered a good explanation depending on the context in which it is requested. This method recognizes the social environment in which explanation questions are asked, and may unify many different existing accounts.

Overview:

In the following dissertation, I will defend an account of algorithmic explanation, which clarifies how they ought to be defined, and gives normative guidelines for making less ambiguous requests for explanations. **In Part I**, I will motivate why explanations are needed in some cases. Without making a strong claim about the specific circumstances where someone ought to be given an explanation for an algorithmic decision, I will make a weaker argument– that it seems there are some cases, sometimes where an explanation is desirable. Then there ought to be a definition that attempts to unify the vast array of different possible explanations. I will consider a variety of question/answer pairs that I've devised to show the wide range of explanations that may be requested about an algorithmic decision. After motivating the need for a more general account of algorithmic explanation that can accommodate my examples, **in Part II,** I will turn to examine a variety of different accounts of algorithmic explanation that can be seen in the literature on Explainable AI (XAI). Crucially, I think that there is not enough attention paid to the philosophy literature on scientific explanation. I will examine some accounts from philosophers of science and will see how well they could serve as an account of algorithmic explanation. The verdict is that none of the accounts can accommodate the wide variety of question/answer pairs! I don't claim that they are deeply wrong, just that they are all too narrow in focus and can't serve as a general definition that we're after in the first place. **In Part III,** I will explain Bas van Fraassen's account of Pragmatic Explanation in detail, expanding on his examples, and weaving it with question/answer pairs from earlier. His framing of an explanation as an answer to a question, and putting a focus on the 'context' in which it arises, is very powerful. This can account for the same question getting quite different, but seemingly correct answers. This account of explanation, without much modification, can serve as a general account of algorithmic explanation, by ensuring that the topic, contrast class, and relevance relation are all made clear from the onset. Finally, **in Part IV**, I suggest a framework for formulating why-questions about algorithmic explanations that forces someone to be clear about what they're asking. I claim that in the specific case of pragmatic algorithmic explanation there are four central features– algorithmic system, action task, input, and output– that are needed in specifying the question and narrowing in on some context. I suggest that if we insist that algorithmic explanations be requested in this form, then we have taken a substantial step in clarifying what is being asked.

I. Background and Motivation:

There is a fact of the matter: complicated algorithmic systems– often described as machine learning (ML) or artificial intelligence (AI) – work very well at achieving some constrained goals. Simultaneously, it can be difficult to explain *how* and *why* they succeed. This is not to say that they are intrinsically superior to algorithmic methods which are designed to be interpretable (although they may be!). Instead, the claim is that scientists and engineers can achieve instrumental success when they use quantitative methods where the internals of the algorithm remain somewhat opaque or unknown– often referred to as a "black box" algorithm. Researchers have shown that in recent years, black box algorithms are being used in high-stakes situations, such as loan determinations, hiring, purchasing of securities, bail decisions, and facial recognition (Rudin, 2019). In these cases, it seems intuitively important that we can offer some sort of explanation as to how and why an algorithm came to a decision. If there were a person deciding these rather than a computer, we'd expect there to be an explanation! Or at the very least, we'd know that there exists some rationale, despite it being inaccessible to us.

Broadly, it's easy to imagine scenarios where we really want our complicated and super important decisions made by computers to be understandable to us. If, for example, someone were given a cancer diagnosis by some medical imaging oncology software, and the consequences of the verdict were to begin chemotherapy, there would be a huge number of follow-up questions, namely about the reasoning that led to such a recommendation. Likewise, when it comes to prison sentencing and the tangible effects on human lives, it seems that an explanation may be needed. If this were just done by a doctor or a judge, then they'd be able to offer some sort of explanation, and while it may not be fully satisfactory, at least there is some rationale for one's life-changing in a major way. On the other hand, there are some cases where algorithmic decisions are harmless– when we use a Snapchat filter and it guesses our age, when a test tries to guess someone's hometown from some of their use of slang, when a generative image algorithm makes flowers of a specific shade. There seem to be cases where there is virtually no harm that can result from algorithmic decisions, and there are cases where there is potentially extreme harm done. However, without a coherent way to demarcate them, a general account of explanation would apply to all of these cases, but may only be used for some. And if we are looking at explanations for low-stakes cases, there may be no *good explanation* at all (which is

fine if it really is harmless), whereas, in high-stakes cases, we may require a much better explanation to justify the harm that may result. Without needing to demarcate these cases, we can take aim toward a general account of algorithmic explanation, knowing that it may not always produce equally good explanations. We're interested in an account of algorithmic explanations that can apply to every algorithmic decision.

Taking this further, some authors have framed this as a rights issue. In 2018, when the GDPR went into effect, it explicitly outlined a *right* "to obtain an explanation of the decision reached after such assessment and to challenge the decision" (GDPR), which codified explanations for algorithms as something of legal interest. Further philosophical justification has been argued by Vredenburgh, who claims that people ought to be informed of "rule-based normative explanations and rule-based causal explanations" (2022). This is to protect individuals' agency over factors that lead to decisions that may impact their life. Without being offered explanations, and understanding ways that their behavior could be changed to influence the outcomes, they've been deprived of agency in their life. The instrumental success of these systems alone cannot justify their use in high-stakes situations– they need to be able to offer explanations in tandem. Further, a good account of algorithmic explanation seems like an essential component of AI alignment– the goal of ensuring the AI systems are learning what we intend them to learn and converging on the values that we want (Christian). As AI systems increasingly improve, having an account of explanations may be crucial to evaluating how well the algorithms are achieving the goals we've tasked them with. As we test an AI, it may be the case that it's performing extremely well, but it's crucial to ensure that it's getting to the outputs for the right reason.

So, if we are using algorithms in high-stakes settings, a general account of algorithmic explanation seems helpful. This can be argued in terms of rights, or in terms of being essential for long-term safety.

If someone agrees with the following:

1. Sometimes, algorithmic decision-makers are instrumentally valuable
2. Sometimes, explanations are required for decisions

Then, you'd agree that there are cases where an explanation is valuable. To start, I will keep my definition of "*explanation*" vague, and later it will be refined. Let's say for now that: it matches what you intuit an explanation is and can include things that are more than a mere description of facts. Oftentimes it is posed as an answer to a question. As we talk about "*algorithmic decisions*", I mean the result of an algorithm when it has some connection or consequences to the world; it's more consequential than an output in isolation.

An example to motivate the rest:

Imagine that someone does most of their banking online. They have been given a check for £190 and using their phone, they take a photograph of it to deposit into a bank account. Later, they ask the following question, after the deposit is incorrect:

VQ (Verbal Question): "Why did this check deposit into my bank account as £100 instead of £190?"

This is a reasonable request for an explanation, framed as a verbal question[1]. However, this request for an explanation seems to lack the level of specificity required to get an expected answer and is underdetermined. This can be seen by the variety of explanations (answers) that may be correct, but have little in common with one another. Let's look at some:

A_1 (Answer 1): Because the 9 was written with a large loop and looks like a 0.
A_2: Because this singular trained neural network being used misclassified the second 9.
A_3: Because the computer system has no human-in-the-loop, and when it makes mistakes like this or is unsure, there is no one to fix it.
A_4: Because the dataset that was used to train our algorithm has many 9s in it that look like 0s.
A_5: Because there was a shadow obscuring part of the image, causing its pixel values to resemble a 9 instead of a 0

[1] I take requests for explanations to generally be questions like this. "Explain why it is the case that X" seems equivalent to "Why is it the case that X". Crucially, that "Why" normally implicitly is a request for an explanation for X

A_6: Because engineers chose to use a neural network instead of using a linear model.

A_7: Because management decided to automate mobile banking.

There may be issues with some of these answers, and some may be better than others. But I think it illustrates the wide variety of possible answers that can be offered to the same verbal request for information. None of these are wrong per se, assuming their assumptions are true. Generally, they just seem to deal with completely different features of the world. Here are some contexts where the question may be asked in corresponding to the answers.

C_1(context 1): Intentionality of people running the bank (A_3, A_6, A_7)

C_2: Causal mechanisms within the algorithm (A_2, A_4)

C_3: Events leading up to the photo taken of the check (A_5)

C_4: Events leading up to the person writing the check (A_1)

C_x: …

I don't claim that these lists are exhaustive either in answers or contexts, these are just some ideas that illustrate the different forms this can take. There seems to be a nearly infinite number of possible contexts and answers that you could come up with. Instead, I wish to focus on how VQ could be asked, perhaps with C_1 in mind and then be answered by someone with C_4 in mind. It all hinges on what each party deems to be important. And with a question this vague there are a huge variety of answers that seem true, or at least minimally satisfying. An engineer who works for the company may be interested in looking at C_2 whereas the customer is likely asking about things like C_1. Further, I don't want to rank them; I just wish to say that they're all appropriate answers at some point, and the business of ranking them all is a topic for another time. *Crucially, I don't think that any existing accounts of algorithmic explanation can account for such a wide range of answers offered here.*

In summary, we are compelled to offer a unified account of algorithmic explanation. By posing an example we see a ton of different answers appear, many of which seem distinctly different in character, but still, seem appropriate answers nonetheless. Our account should accommodate all of these. And I will propose an account that does.

II. Existing Accounts of Explanation

We want a framework for understanding algorithmic explanations, that is able to accommodate the wide range of types of questions that can be asked and the seemingly larger variety of answers that are offered in tandem. In this section, I want to look at how this has been attempted thus far. The conclusion will be: that none of these accounts are deeply wrong, but individually they are not broad enough to capture the explanatory gamut. Many of the existing accounts of algorithmic explanation seem to take focus on one specific type of explanation[2]. While such a method may be able to provide one of the sample answers (A_n), it would also unnecessarily disqualify other answers.Many accounts put forth in the machine learning literature don't seem to be giving due reference to the philosophical literature on scientific explanation. While algorithmic explanation has become a recent topic of interest, the more general notion of scientific explanation has been studied for much longer. What connects much of the philosophy on the topic, is the claim that *'explanation'* is a universalizable concept– that all scientific explanations are just a more constrained type of what we intuit to be an explanation for something (Woodward, 2021). As put by James Woodward:

> *"On a methodological level, we should expect causal explanations in different areas of science to share at least some structural features with causal explanations in more ordinary contexts. We should see scientists who construct such explanations as attempting to satisfy some of the same explanatory goals and interests as people who construct explanations in ordinary life, but as achieving those goals by appealing to knowledge that is more rigorous, detailed, and systematic. Explanation in science does not give us something that is fundamentally different in kind from explanation in more ordinary contexts, but rather, as it were, a better version of the latter." (Woodward, 2001, p. 20)*

I find this a compelling view– that explanations of all types have some commonality, and that scientific explanations are the product of increasing our standards for what is needed in an explanation. Then, in the list of answers to the verbal question, some may be scientific

[2]It would be in bad faith to say that these specific authors are holding their specific tool or definition of an explanation in one context to be a general account that applies to all possible acceptable explanations.

explanations (internal software mechanisms)[3], while others (intentionality of agents) can be considered perhaps a less constrained, general explanation. Scientific explanation is a subset of the explanation, which has the same types of goals– helping humans to understand– and shares some structural similarities. Further, I think we can take this unified explanation idea offered by many philosophers without too much baggage– we can remain agnostic about the ontological status of explanations, as long as we believe that this description seems to capture the essence of what we're after. If, after further discussion, we have an account of what it means to be an explanation, and this works for nearly all of the examples we're thinking about, then we've probably found some emergent structure of what it means to be an explanation, regardless of its deeper fundamental characterization.

a. Accounts of Explanation in ML Literature

In the computing literature, there has been recent interest in designing systems that are interpretable, so that explanations can be easily extracted. However, as pointed out by Lipton (2018), even the idea of interpretability is underspecified. Oftentimes, explainable machine learning is viewed as a tool that needs to be constructed which gives someone information that serves as an explanation. Attempt to further specify and characterize the forms this could possibly take have been done in many ways, and one such way is to frame it as either a 'post hoc' or 'transparent' explanation. Still, this categorization by itself is vague and doesn't easily account for the variety of solutions that are offered, but is an acceptable way of thinking about it. Let's consider a system that is being modeled by a directed graph below:

```
Input → Model → Output
```

If we use this to consider the example of our handwritten check in VQ, then: the input is a picture of a single handwritten digit, and the model is an undefined classification algorithm, and the output is the category is 0 (when it should have been 9). Post hoc explanations look only at the input and output. There may be some benefit to knowing vague properties of the model, as shown by Artelt (2019) which develops specific methods to better look at counterfactuals, but

[3]I'm unclear about if this is true (especially when put with a Woodward quote) because he claims that matters of logic or mathematics aren't explained– they are simply described. But this is a minor detail that I will ignore.

generally, example-based models skip over the details of the model. These can also be referred to as 'post hoc' explanations. Other types of explanations seek to use the model part of the graph, not just to aid in the analysis of inputs and outputs, but as something intrinsically important. They may try to explain the steps that lead to an explanation along the way.

i. 'post hoc'/ 'example'

With post hoc methods, it is assumed that the causal intricacies of the model cannot provide us with what we need– we ignore the model part of the graph altogether. Instead, we look at the outputs that it gives us, and seek an after-the-fact way of explaining how the output came to be. In these scenarios, the model is often referred to as a 'black box', meaning that it is impossible to see what is happening inside, and how the computation is arriving at a solution. While some accounts may object to this characterization– there is nothing fundamentally unknown about the computational steps that are occurring, rather the difficulty is in connecting component steps to any sort of coherent narrative or explanation– with this assumption we can still try to explain. One category of this is a counterfactual explanation, which can be loosely defined as:

> *In order to explain why O_1 was returned (as opposed to O_2) from input J, you can be offered an input J_1 that is similar to J, with some parts minimally changed, that would return O_2.*

The counterfactual account seems to have a few basic strengths. The first is that computers are very good at evaluating counterfactuals, in the sense that you can create any input, and see what the output is. You can say, with utmost confidence, what the result of this image classification task would be if a region of the image was dark, for example. This is a promising account because we can know the truth value of a counterfactual. Asking a human to try to explain by counterfactual reasoning seems methodologically unsound, whereas a computer can do this immensely well. There is debate about the correct formalism for counterfactuals (Hudetz), specifically, about when a statement is actually true, but in this discussion, we can float above that level of detail. In the ML literature (Verma), there are desirable features of counterfactual explanations that easily connect language to the mathematics of algorithms. There

are countless ways to frame the desirable features of counterfactuals and this is just one such way. Crucially, any specific algorithmic implementation of counterfactuals used for explanations ought to adhere to a logical system that is good, but inevitably, some arbitrariness will enter into the equation.

First, is the notion of validity, that if we change the input in such a way, it does in fact produce the output. Next is actionability– the changes that are suggested are things that can be controlled, as opposed to fixed things like gender, age, or race; suggesting that someone change races to get approved for a loan may be useful information if you are checking an algorithm for bias but is not helpful to an individual applying for a loan, and if a counterfactual explanation is requested for the purposes of getting a loan, it is important that it is in fact actionable. Next, it is desirable for the counterfactual to be sparse: that only a few of the input categories should be changed, as opposed to changing all of them in a slight way. The idea is that it becomes more understandable and actionable for a person. Perhaps, knowing that getting a greater income will help loan approval, and this is much better than slight changes to all of the following things: zip codes, income, career, and education. Causality is the last, which means that any inputs that depend on other inputs ought not to be changed in isolation. A well-formed logic of counterfactuals ought to enforce these features, whereas things like Interventionist Semantics may not capture things like causality conditions, for example. Like any such list of concerns in an applied science paper, it is surely not exhaustive, and there is some overlap. Rather than viewing this as some rigorous attempt at providing a universal account of how desirable counterfactuals ought to be constructed, I prefer to view it as a (likely pretty good!) guideline for evaluating different implementations of counterfactuals and whether it produces information that is telling. Further, Verma has compared 39 different specific implementations, that are meant to work with varying algorithms that may be of instrumental value. The guidelines serve as a way of loosely ranking specific methods by identifying features that we may care about.

The general issue I have with counterfactual explanations is: that while they can tell you a lot about the performance of a specific algorithm, they can't tell anything about how the algorithm came to resemble its current form. You could use a series of generated counterfactuals to get an idea of how a classifier is working (for example, how many extra dark pixels can you add inside the circle of the 9 before it is misclassified), and this may give you some pretty useful information about a classifier and how it behaves. But, I think this is not sufficient for all of the

explanations we're considering. First, only a few counterfactuals can be computed. The handwritten dataset we've been referencing is made of small 8x8 pixel images. If we decided that we were restricting the pixel values to 8 shades from white to black, there would be 8^{8*8} possible input images to check all of the counterfactuals[4]. What I'm describing here is the maximal amount of information that you could possibly get with this counterfactual framework, and that checking every possible output from every possible input is impossible. Even if this weren't impossible, it would still not answer some questions that we may have.

Let's imagine a super simple example of a loan classifier. The only inputs are the applicant's income and the amount they request. If they are requesting more than 5x what they earn in income, they will be declined. It could be visualized as follows:

```
If loan ≥ (5 * annual income):      accept
Otherwise:                          decline
```

Unlike our image examples, we can see all the possible inputs and the corresponding outputs. There is no mystery at all as to what happens when we apply for a loan. This is an idealized version of a tool for counterfactual explanation. The table below looks at two possible counterfactual explanations for the declination of the loan, and how changing one's income or the loan amount requested may change the outcome.

Income	Loan	Result
£10,000	£60,000	decline
£10,000	£50,000	accept
£12,000	£60,000	accept

[4]An overwhelming majority of these pictures will just look like static, and there would be a much smaller subset that are actually interesting to us. Additionally, there are strong assumptions here that there are NO monotonic relationships that we can use to simplify things. But 8^(8*8) has over 55 zeros at the end of it, and this is perhaps the simplest image example that can be concocted. If we thought of a black and white image, with 8-bit color precision, and is 1080p resolution, there are 256^ (1920*1080) pixels. This number has 5 million zeros. We are talking about 60,000 times the number of atoms in the observable universe. It's completely absurd to imagine how many different pictures there can be. Let's just say: it is for all intents and purposes impossible to evaluate everything counterfactually in these examples.

For this example, we have full counterfactual knowledge, but we clearly can't get explanations for all of the questions we may have. Crucially, we have no idea why the number five is chosen as the multiplier here. Or why it is a simple binary classifier– maybe there should be a third option, where it gets manually reviewed. Or why there is not more information requested from applicants. Or why a potentially trivial amount of money– say an income difference of a few hundred dollars– could make the difference. This supports the idea that there are explanations we could request about this algorithm that cannot be answered with full counterfactual knowledge. To generalize these concerns to other examples, in the image case, you can input a variety of different pictures of 9s and 0s and have it tell you how it would evaluate them. You'd get some idea why it was classifying things, but it can't tell you how this exact algorithm came to occupy the role that it is currently filling. Someone could reasonably ask questions like: How was this model trained, and with what information? How was the model selected? Why were these image dimensions chosen? Why was this level of accuracy deemed appropriate? These are situations that still fall into algorithmic explanation, but are unrelated to checking how input changes map to output changes.

Perhaps we're interested in something that could be described as world construction– when our input goes into an algorithm it is subject to 'laws of nature' in a sense, and by running a series of experiments and seeing how it behaves, we might be able to make claims about emergent properties. But no amount of testing inputs against this framework will tell us *how* or *why* it came to have the laws and relations that it does. In the physical world, these are questions that seem to capture people's attention because they're fun to ponder, and not because we expect to easily find satisfying answers to them. But when we think about our graph, which someone is responsible for creating, the pseudo-existential questions like "Why did you use 5 as a multiplier?" is at the very least something we might want to ask, and seem squarely included in what we're trying to define.

In conclusion, I've made the claim that counterfactual explanations offered in the computing literature cannot be a general account of algorithmic explanations. But, I think they are an incredibly promising part of algorithm explanation. A general account of algorithmic explanation then should recognize that sometimes these are good explanations. In the example loan case, counterfactual information might be something that the applicant is satisfied to hear,

as long as they were asking for this type of information, and everyone involved had a mutual understanding of the context.

ii. 'transparent'

Another framework for explanations is to look into the model part of the directed graph that we were ignoring in post hoc explanations. This can be done in a few ways: 1) by designing a simple model on purpose 2) by trying to 'break open' the black box, and get an idea of what is happening inside the model, 3) by creating some new model that captures the properties of the more complex model, or in other ways.

This is a delicate issue. There are some authors who altogether reject this type of reasoning, claiming that sufficiently complex models are hopelessly opaque, and cannot be explained, as argued by Rudin (2019). In another paper, Rudin demonstrates that there can be cases of immense success when you are explicitly optimizing for simple, interpretable models and using machine learning to find a model that works best (2019). For example, when determining recidivism risk, the authors were able to prove that a simple verbal algorithm is as accurate as state-of-the-art machine learning models, and is much easier for everyone involved to understand what is causing someone to be considered at risk of reoffending. It took the form of:

> *"if the person has either >3 prior crimes, or is 18–20 years old and male, or is 21–23 years old and has two or three prior crimes, they are predicted to be rearrested within two years from their evaluation, and otherwise not" (Rudin, 2019)*

So in these cases, perhaps we should not be designing models that are opaque or at least not when we are seeking explanations– the harmless cases of algorithmic decisions previously mentioned would be okay to design in an opaque way. But this strong claim, that we are incapable of providing satisfactory explanations for a large class of models that we've shown are very successful, doesn't seem to be a general truth. An account of explanation should be adaptable to possible changes: a breakthrough in the ability to construct tools to estimate complicated models, or maybe some explanations that are understandable only to subject matter experts. This is not to say that either of these would be good explanations for everyone, but they seem to be closer to explanations than not.

To a researcher, a statement like “This 9 was misclassified as a 0 because the training dataset has a lot of 9s with big loops” may actually be the type of answer they’re looking for. It’s not a justification for why someone lost $90 from their check, since presumably, the algorithmic explanation is just a description of *part* of the reason that their bank deposited the wrong amount. Rudin is right to say that in high-stakes cases we ought to have a system that can offer us more than a minimally acceptable explanation– a good explanation, and the type offered for the handwritten digits just isn’t satisfactory when the consequences are years of a person’s life. But when the consequences are having to make a phone call to get your bank to fix their mistake, it becomes less clear. This seems to be an issue with how good an explanation is, rather than an explanation being minimally satisfactory. In cases where there is a lot at stake, we may want explanations of a certain form, that are very telling, so that we have a moral justification for causing big changes in people's lives. Then, there seems to be a relationship between the strength of an explanation, and what we can use it to justify. I think this is an interesting secondary question, which I will not deal with– crucially it involves the act of ranking explanations and connecting explanations to justifying harm to others.

To close, I think this literature fails as a general account of explanation because it tends to dismiss weak explanations because they aren’t strong enough to justify certain types of actions, but many of them are still explanations nonetheless! It seems true that explanations are not equally telling, but a general account of explanation will capture the good as well as the bad.

Other ways of trying to create transparent models involve technology like salience maps which are useful in making sure that a model is picking up the right features (Ribeiro). In this case, it ensures that when working with images we can see what parts of the image it is focusing on. Similar technology found explanatory use, as explained by Christian when a commercial team was able to predict the age and sex of a patient from an image of their retina (2021, p. 107). Salience maps were able to verify in the image where they were getting these clues from– blood vessels for age and the optic disk for sex. In this case, the authors are modeling a simplified version of the algorithm (and using quite a few counterfactuals) to attempt to understand what is happening in the ‘black box’. However, these models are necessarily making huge estimates, and offer nothing more than a loose idea of what is happening inside.

b. Accounts of Explanation in Philosophy

The literature in philosophy of science seems to agree that explanations are more than a mere accurate description of something that is happening– they are a statement about why something is happening. I will take a very brief look at a few accounts, trying to connect their essence to the accounts found in computing literature. Expanding on the introduction earlier, there is an assumption that explanations of all types have some commonality, as said in Stanford's Philosophy of Encyclopedia.

> *"Much of the recent philosophical literature assumes that there is substantial continuity between explanations found in science and some forms of explanation found in ordinary, non-scientific contexts. It is further assumed that it is the task of a theory of explanation to capture what is common to both scientific and some ordinary, non-scientific forms of explanation." (Woodward, 2021)*

Many of the approaches take desirable and undesirable features of explanation that seem intuitively true and compare accounts to match these. Relying on examples where people tend to have a strong intuition, it's easy to illustrate the various features of explanation.

Philosophers seem to agree that a good account of explanation will do the following: avoid explaining by referencing facts that are nothing more than accidental generalizations, correctly identify cases of explanatory asymmetry, not include things that are irrelevant, avoid explanations that reference effects when they should be referencing common causes, have some predictive use, and be able to explain unlikely things when they do happen. In the case of algorithmic explanations, these all seem desirable[5]. Now, we will examine different accounts of explanation, knowing that these are the types of features they are trying to capture.

i. Statistical Relevance

Statistical Relevance (SR) is a simple account of explanation that only looks at the probability of events, introduced by Wesley Salmon. He would say that C explains B if the probability of C given B is higher than the probability of C alone. This supposes that an

[5] There's a question of whether or not the low P criteria is needed since most of our algorithms are deterministic. While there may be no true randomness, we still are justified in modeling some of our systems as random.

explanation is not an argument, or a logical system, just a collection of information that is relevant to an event. Formally:

> *"Given some class or population A, an attribute C will be statistically relevant to another attribute B if and only if $P(B|A.C) \neq P(B|A)$—that is, if and only if the probability of B conditional on A and C is different from the probability of B conditional on A alone." (Woodward, 2021)*

There are three issues that can be pointed out here. First is the assumption that what is being explained eventually reduces down to completely indeterministic things– that probability is occupying some deep and true ontological status. This is not the case for many of our algorithms, which behave in completely predictable ways. Further, as a practical issue, this is only a useful general account if we have a great knowledge of various probabilities in our system. Last, it has been logically shown that it is possible for causal relationships to not be captured by the SR model! This is the big issue. If we can't capture a known causal relationship, we must dismiss this account.

ii. Causal Mechanisms/Counterfactual

Some of the more promising explanation types are counterfactual, like James Woodward's writings of causal explanations (2003). I believe these face the same issues that the implementations in computer science face. Namely, with computers, even knowing all of the counterfactuals of our system doesn't let us know how it came to be. Then, a more general account of counterfactual explanation could ask things like: "What if we had trained this model with a different dataset?", and try to explain other things about the model that can't be answered by testing it a bunch of times. This is potentially a big step in the right direction.

In the computing literature, there don't seem to be accounts that do this. If we think about some of our bigger algorithms as complex systems, then there are further questions about how we can trace a metaphorical 'mark' through the system. Even if we set this problem aside, dealing with the full range of possible counterfactuals is an entirely different issue. Everything that involves running inputs and seeing outputs seems well studied, or at least promising at delivering types of explanations that we may be interested in, but if we start asking questions

where we expect an answer like A_7 (managements intention to automate a task), then it is difficult to say we know counterfactual truths about what would have happened if they hadn't automated it. To make an account of explanation from the causal model, the CS literature would need to move outside the realm of the model.

iii. Laws of Nature

This is worth mentioning only because it is the most famous account of explanation– that explanations are in fact arguments, either deductive or inductive, that take as their inputs laws of nature and various other assumptions (Hempel). There doesn't seem to be any sort of analog in the world of algorithmic explanations unless we think of algorithms as worlds in and of themselves, which I alluded to earlier as a possibility that I am ignoring.

c. Conclusion

This has been a long and winding section, which looked at 1) what we think explanation ought to look like, 2) how computing researchers have approached it and 3) how philosophers have tried to tackle the problem. We haven't yet seen an account that can handle the vast array of different explanations that are needed. Now, we will.

III. Pragmatic Account

Bas van Fraassen offers an account of pragmatic explanation in his book *The Scientific Image* (1980), where he first discusses the shortcomings of other accounts before making clear his novel pragmatic account. Crucially, previous accounts of explanation thought that theory and fact alone could give explanations, and ignored the broader context of the situation. Succinctly:

> *"The discussion of explanation went wrong at the very beginning when explanation was conceived of as a relation like description: a relation between a theory and a fact. Really, it is a three-term relation between theory, fact, and context. No wonder that no single relation between theory and fact ever managed to fit more than a few examples!" (van Fraassen, 1980)*

To reconstruct his account, an explanation is an answer to a why-question, so a theory of explanation is a theory of offering answers to why-questions. It is important that explanations can first be confirmed as minimally acceptable answers before being compared to one another. First, it needs to be made clear what is being asked– namely, what is to be explained, and there is immense context dependence here. Then, answers ought to be evaluated for their quality, as opposed to their minimum adequacy. I will reconstruct the terminology and definitions that van Fraassen uses and elaborate on one of his examples.

> ***Question Definition****: A question Q = <P_k, X, R> is a triplet of a topic (P_k), a contrast class (X = {P_1, P_2,... P_k..., P_n}), and a relevance relation (R).*

Every well-posed question will be describable in this way; if certain parts of the context are unclear, there will be multiple Q triplets that can arise from the same verbal question. The topic is the true statement, the thing that is trying to be explained, as opposed to the rest of X which are propositions that are not true. The answer bears relation R to <P_k, X>. van Fraassen resists a strict and formal definition of R, but shows that it can take varying forms of specificity– from statements like "give me a statistically relevant preceding event not screened off by other events" to "give me a strong motive".

To expand on his example, say that someone is at a table that has 10 long conducting wires placed upon its surface, and they believe them to all be identical in shape, material, temperature, and other relevant features. Subsequently, one of the wires warps in shape, so we seek an explanation by asking:

> Verbal Question[6]: "Why did this conductor warp?"

Further, let's assume in this context that the person is asking: "why did conductor 1 warp, as opposed to 2-9?". If this is true, then we can reformulate their question in van Fraassen's form seen above.

---

[6] van Fraassen doesn't define it this way, but it improves clarity.

P_k (topic) = "conductor 1 is warped"[7]

X (contrast class) = {"conductor 1 is warped", "conductor 2 is warped",... "conductor 9 is warped"}

R_1(relevance relation) = explain the electromagnetic fields around the conductor

R_2 = explain the intentionality of people who could control the E&M field

Now that we have a why-question seeking an explanation, we can begin to offer an answer.

Answer Definition: *P_k in contrast to the rest of X, because A.*

Then here are some propositions that could be offered in place of A to serve as an answer to the Questions above, where A bears relation R to $<P_k, X>$.

A_1: There was a strong local magnetic field applied underneath the table below 1, which did not extend far enough to affect 2,3…10.

A_2: Bulut told Nala that he wanted to warp conductor 1, and he likes to keep his word.

In this scenario, the same verbal question "Why did this conductor warp" can be viewed as two different questions, when formalized as in this account, due to the different R. Then an answer to each is different since A_1 and A_2 are related by R to the question which is different. The same verbal question becomes different why-questions, and since why-questions are different requests for explanations they will have different answers. In one scenario, someone may be asking for a scientific explanation of why exactly the conductor warped (perhaps they are a student of physics). In another, someone may be asking why one intended for a conductor to be warped (and takes the physics of warping a wire to be unimportant). In these cases, the contrast class may very well be the same, while the relevance relation, determined by the context, can lead to the question being interpreted and answered in two completely different ways. So, in our example, the contrast class is the same between the two cases.

[7] When asking a question: "why did conductor X warp?", vF claims that "conductor X is warped" is a presupposition of the question, and we take the proposition to be true.

However, it is also possible to have the same verbal question correspond to different contrast classes and relevance relations. Take for example:

P_k (topic) = "conductor 1 is warped"
X_1(contrast class) = {"conductor 1 is warped", "conductor 1 not warped"}
R_3(relevance relation) = explain the tensile strength of the material and how a force overcame this to bend it

The person may be aware of some electromagnetic field that is effecting conductor 1, may not be concerned about the other wires, and is instead wondering how the force was able to materially bend the wire.

The same verbal question leads to a variety of formal questions, the set of: $\{Q_1 = \langle P_k, X, R_1\rangle, Q_2 = \langle P_k, X, R_2\rangle, Q_3 = \langle P_k, X_1, R_3\rangle ...\}$, where some may share the same contrast classes, and some share the same relevance relations. There are many questions! Now, to finish the reconstruction of van Fraassen's account, a direct answer is defined as follows:

Direct Answer Definition*: B is a direct answer to question $Q = \langle P_k, X, R\rangle$ exactly if there is some proposition A such that A bears relation R to $\langle P_k, X\rangle$ and B is the proposition which is true exactly if the following are true:*

1. *P_k is true*
2. *For all $i \neq k$, P_i is not true*
3. *A is true*

This is just to clearly enumerate some of the intuition discussed above. Now, let's return to our old example. Once again we are wondering "Why did a check deposit into my bank account as £100 instead of £190?". It's easy to see the topic and contrast class here, but the relevance relation is what has been left wide open to interpretation based on context.

P_k (topic): "check deposited as $100"
X (contrast class): {"check deposited as $100", "check deposited as $190"}

$\underline{R_1}$: explain the intentions of people running the bank

$\underline{R_2}$: explain some causal mechanism within the algorithm

$\underline{R_3}$: explain some events leading up to the photo taken of the check

$\underline{R_4}$: explain some events leading up to the person writing the check

Then, the true propositions A, which would be offered as an explanation, vary by their relation to $<P_k, X>$, and we can get different answers like:

$\underline{A_1}$: management decided to stop using humans to review mobile banking deposits, and approved an electronic system with a 0.01% error as acceptable

$\underline{A_2}$: a specific architecture of neural network was used and trained on the MNIST dataset, the model trained and minimized error in such a way that when the photographed 9 is used as an input, it becomes misclassified

$\underline{A_3}$: the photo of the check that was deposited was taken with a shadow obscuring one of the letters, which caused its pixel values to resemble a 9 instead of a 0

$\underline{A_4}$: Roman, who wrote the check, draws his 9s with a very large circle, such that they resemble 0s sometimes.

These are the same types of questions, contexts, and answers that we posed in the introduction. I believe these are all adequate answers to the verbal question– as long as the person asking for an explanation expects this type of answer in response. Crucially, this theory of pragmatic explanation takes all of these to be serious, correct answers to the same verbal question, by interpreting it as a number of different questions. As we've seen, other accounts of explanation may disqualify some of these responses for the sake of offering a neat definition. I think the key to algorithmic explanation that has been missing is forcing the question to be clear about what it is asking. If a question is underdetermined by the context, then relevance relation must be assumed to offer an answer that is satisfactory, but it will only be satisfactory for one formal question where $Q = <P_k, X, R>$. In this case, we may be answering a different question than the question-asker expects us to answer; this will never lead to clarity! When the same statement can be viewed as a near-infinite number of Q triplets, our confusion arises.

Then, we ought to be concerned with how we can make clear what question we are asking and answering. By this account, much of the issue of algorithmic explanation is rooted in vague questions, where the asker and the answerer interpret them differently. Further, as we examine different contrast classes and relevance relations, we can see that this can accommodate the other accounts of explanations thus far. If we are wondering why our check has been deposited wrong, and in our mind are thinking something like R_1, then an answer like A_1 would be what we're looking for. If most of the issues with algorithmic explanations come from unclear questions, it would be good to have guidelines about what is important to improve clarity.

IV. Pragmatic Guidelines for Algorithmic Explanations

Thus far, I have restated van Fraassen's theory with some examples and shown how it captures the variety of answers to a verbal question requesting an algorithmic explanation. Further, it explains why explanations themselves can be so broad– each situation's relevance relation allows for different acceptable answers. Now, I wish to offer a formulation of why-questions for algorithmic explanations that force the asker to be clear in what they are actually asking. This is a normative suggestion, rather than a claim of deep truth.

Questions related to algorithmic explanation can and ought to be phrased in the following way:

Clear Question Definition*: Why has (W) (action X) (Y) (Z)?*

W: algorithmic system

X: task of an algorithm

Y: input

Z: output

Where **W** is the type of computing system being used (linear regression, neural network, decision tree, etc...). In a maximally specific case, this would be a single instance of a trained model, a set of fully defined parameters, or some other narrow definition. This could also be defined more vaguely, but explanation clarity will suffer. Further, this would include low-level considerations. Authors like Creel (2020) have captured the immensely wide scope of all the

different points that opacity can enter: from the algorithmic theory, the specific coding, and the implementation in hardware.

Next, action **X** can be thought of as something that does the following: classify, segment, project, write, illustrate, or approve a loan. It is the task that relates Y to Z, and the action that our algorithm has been tasked with doing. It is worth keeping this separate from X, since the same algorithm may have a variety of output tasks that can be selected and obtained from it[8]**.**

Y is the input to our system, which may be a single image of a handwritten digit, an entire category of numbers, or a loan application, for example. This is what we are providing our system, and is necessary to produce Z, the output.

Z is the result from the system– maybe this is a classification label (which says that our handwritten 9 is actually a 0), or the verdict of a loan application. A lot of the questions may be framed around why a particular Z was selected, as opposed to another Z.

I think these are all distinct and needed, and help transform a verbal question into a pragmatic question, which we've shown can have an answer. Crucially then, someone ought to be clear about which portion of the question the contrast class (CC) is arising from, and what sort of relevance relation (R) they are looking for. I'll remain agnostic about if these four considerations are sufficient; rather, I just think these properties emerge as being important in the case of algorithmic decisions and I see no harm in being specific about what they are since there is oftentimes a fact of the matter. Further, I don't see how specifying these would make it less clear.

So, for the last time, let's return to the example. The verbal question "Why did a check deposit into my bank account as £100 instead of £190?" may become: Why has (this specific neural network) (classified into integer category) (this image of a handwritten 0) (integer category 9)? While the grammar has suffered, the request for an explanation is more specific. It forces people to make clear what exactly they mean when they ask for an explanation, offers an easy way to determine which of the four is where the contrast class deviates from, and becomes a good starting point for making clear a relevance relation.

---

[8] This can be seen in Multitask Learning (where there are potentially a number of different tasks that are being done, but we may only be concerned with one), or in cases of classical statistical methods, where we can use the same model in regression or classification. The same type of model may be able to do different things, so making clear the action X is a good step in reducing ambiguity.

First, this can help specify which contrast class is the one of interest. Let's examine all of the different forms this could take.

Contrast Class W:

- Why has **(Neural Network)** (classified into integer category) (this image of a handwritten 0) (category 9)?
- Why has **(linear regression of handpicked features)** (classified into integer category) (this image of a handwritten 0) (category 9)?
- Why has **(human reviewer)** (classified into integer category) (this image of a handwritten 0) (category 9)?

Contrast Class action X:

- Why has (Neural Network) **(classified into integer category)** (this image of a handwritten 0) (category 9)?
- Why has (Neural Network) **(classified into integer category only if confidence >90%)** (this image of a handwritten 0) (category 9)?

Contrast Class Y:

- Why has (Neural Network) (classified into integer category) **(this image of a handwritten 0)** (category 9)?
- Why has (Neural Network) (classified into integer category) **(different image of a handwritten 0)** (category 9)?

Contrast Class Z:

- Why has (Neural Network) (classified into integer category) (this image of a handwritten 0) **(category 9)?**
- Why has (Neural Network) (classified into integer category) (this image of a handwritten 0) **(category 0)?**

Contrast Class W is asking a question about why this particular algorithm is being used, as opposed to some other classifier type. Contrast Class action X is asking why the action is being taken with the parameters it has. Contrast Class Y is asking why this particular case failed, as opposed to another similar digit. Contrast Class Z is asking why the digit has been miscategorized, which is likely what the majority of questions about this one are asking about. Questions like these all seem important!

For all of these cases, we can further specify relevance relations, where we are seeking some counterfactual connection, some agent-intentionality reason, or something else. By pointing out these features that seem common to cases of algorithmic explanation, we then force the question-asker to be clear about what they are asking! First, we are making clear that the question is "Why P_k as opposed to the rest of X". Thereafter, we need to clarify R. R, in this case, is what kinds of reason do you think are reasonable that P_k is chosen rather than the rest of X, and

this could be something like “explain with input-output counterfactuals”, or “explain with reference to the management's intentions”. Without making this clear, we will run into some of the same issues as before.

While the definition of R was left a bit vague by van Fraassen, I think this is where it connects to some of the narrow and specific accounts of explanation discussed earlier. R could take the form of “explain with a counterfactual that meets criteria in Artelt’s 2019 paper”, or “explain with a local model salience map”, or “explain with reference to human intentions”. *Crucially, if we take a specific explanation variety to be satisfying in some contexts, and we specify this as R, then we’ve potentially unified the other accounts of explanation into our pragmatic account.* The critics of this account of scientific explanation claim that without more constraints about what can be a relevance relation, nearly everything can count as an explanation. Perhaps there are more constraints that should be waged at this point. But in algorithmic cases, some of these narrow specific accounts of explanations seem to be a satisfying R, and determining which kinds of explanations are better and worse is a follow-up question.

So, my recommendation when asking for an algorithmic explanation is:

1. Specify the algorithmic system, task, input, and output
2. Emphasize which of them forms the contrast class
3. Be clear about what kind of relationship you want between event and explanation

If these steps are taken, I think that the wide range of examples throughout the paper are accommodated, and we have a holistic way of approaching explanations when computers are responsible for consequential tasks.

V. Conclusion

In summary, over the course of this dissertation, I’ve motivated that, sometimes– whether it be due to legal obligation, due to rights-related obligations, or otherwise– there is a need for explanations of algorithmic decisions. We need to define what it means to ask for an explanation. Existing accounts offered by computer scientists don’t seem to capture the range of answers that can be offered, and if we are drawn to the idea that explanations in science are highly related to explanations in everyday life, then we believe there should be an account that unifies both. By

reconstructing Bas van Fraassen's account of Pragmatic Explanation, we've seen that it captures the explanatory gamut by putting central importance on the notion of the context in which the question is asked. Taking this pragmatic account to be desirable, our goal shifts to ensuring that the explanations we request for algorithms are clear. By specifying four emergent features—inputs, outputs, tasks, and algorithm– and how we want the explanation to relate to the decision, we take a step towards a direct correspondence from verbal questions to formal questions. A majority of the confusion arises from hopelessly undetermined verbal questions, where there is a variety of true answers that the asker does not want. Only after we are clear about the form of an adequate explanation can we ask how good it is and what it can be used to justify. I hope that I've taken a small step forward in clarifying the prerequisite question: what is an algorithmic explanation?

Word Count: 9,509

Citations:

Artelt, André, and Barbara Hammer. "On the computation of counterfactual explanations--A survey." arXiv preprint arXiv:1911.07749 (2019).

Christian, Brian. “Transparency.” *The Alignment Problem: Machine Learning and Human Values*, W.W. Norton & Company, New York, 2021.

Creel, Kathleen A. "Transparency in complex computational systems." Philosophy of Science 87.4 (2020): 568-589.

“GDPR Archives.” *GDPR.eu*, https://gdpr.eu/tag/gdpr/.

Hempel, Carl G. *Aspects of Scientific Explanation: and Other Essays in the Philosophy of Science*. Free Press, 1970.

Hudetz, Laurenz, and Neil Crawford. “Variation Semantics: When Counterfactuals in Explanations of Algorithmic Decisions Are True. [Preprint].”

Lipton, Zachary C. "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue 16.3 (2018): 31-57.

Psillos, Stathis. *Causation and Explanation*. Acumen, 2002.

Ribeiro, Marco Tulio, et al. “‘Why Should I Trust You?".” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016

Rudin, Cynthia, and Joanna Radin. "Why are we using black box models in AI when we don’t need to? A lesson from an explainable AI competition." (2019).

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1.5 (2019): 206-215.

Van Fraassen, Bas C. *The Scientific Image*. Clarendon, 1980.

Van Fraassen, Bas C. “The Pragmatics of Explanation.” *American Philosophical Quarterly (Oxford)*, vol. 14, no. 2, 1977, pp. 143–150.

Verma, Sahil, John Dickerson, and Keegan Hines. "Counterfactual explanations for machine learning: A review." arXiv preprint arXiv:2010.10596 (2020).

Vredenburgh, Kate. “The Right to Explanation.” *The Journal of Political Philosophy*, vol. 30, no. 2, 2022, pp. 209–229., https://doi.org/10.1111/jopp.12262.

Woodward, James. *Making Things Happen a Theory of Causal Explanation*. Oxford University Press, 2003.

Woodward, James. “Scientific Explanation.” *Stanford Encyclopedia of Philosophy*, 2021.