Robustness in Machine Learning Explanations: Does It Matter?

Leif Hancox-Li
leif.hancox-li@capitalone.com
Capital One
New York, New York, USA

ABSTRACT

The explainable AI literature contains multiple notions of what an explanation is and what desiderata explanations should satisfy. One implicit source of disagreement is how far the explanations should reflect real patterns in the data or the world. This disagreement underlies debates about other desiderata, such as how robust explanations are to slight perturbations in the input data. I argue that robustness is desirable to the extent that we're concerned about finding real patterns in the world. The import of real patterns differs according to the problem context. In some contexts, non-robust explanations can constitute a moral hazard. By being clear about the extent to which we care about capturing real patterns, we can also determine whether the Rashomon Effect is a boon or a bane.

KEYWORDS

explanation, philosophy, epistemology, machine learning, objectivity, robustness, artificial intelligence, methodology, ethics

ACM Reference Format:
Leif Hancox-Li. 2020. Robustness in Machine Learning Explanations: Does It Matter?. In *Conference on Fairness, Accountability, and Transparency (FAT* '20), January 27–30, 2020, Barcelona, Spain.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3351095.3372836

1 INTRODUCTION

Explanations of how machine learning (ML) models work are part of having accountable and transparent machine learning. But what counts as a good explanation in machine learning? The machine learning research community has come up with many methods to produce explanations, but there is little explicit discussion of desiderata for machine learning explanations, even as these desiderata are appealed to when evaluating particular explanation methods.

One difficulty is that ML explanations are used for diverse purposes, audiences, and models. Different purposes will lead to different desiderata. Certain types of explanations are inappropriate for non-technical audiences, for example. For this reason, trying to argue for a single set of desiderata for all ML explanations is likely to be a fool's errand.

Nonetheless, I will argue for the merits of one particular desideratum that has not been explicitly discussed, but which I think applies across many purposes, application contexts, audiences and models. This is the desideratum of a particular kind of objectivity: the degree to which an ML explanation sheds light on patterns in the world.[1] Objectivity is threatened when there are multiple candidate explanations that are equal or almost equal on other desiderata, but only one of those candidate explanations is presented as the correct explanation.

A related discussion about ML explanations concerns the importance of robustness (or stability, as it's sometimes called) [26]. Alvarez-Melis and Jaakkola make a brief remark towards the end of [2] that robust explanations may be more important if the goal is to understand not just the model, but the underlying phenomenon being modeled. I flesh out this remark by drawing on a long tradition of both scientists and philosophers who have argued that robustness is an indicator of reality. The epistemic advantages of robustness that they describe, I argue, extend to ML explanations: robust ML explanations are desirable for the same reasons.

After showing that objectivity has been an implicit desideratum for some AI researchers, I provide both epistemic and ethical reasons for seeking objective explanations. The epistemic reasons apply not just when we want to find out about real patterns in the world, but also when we're trying to improve a purely predictive model. The ethical reasons I raise dovetail with worries about the arbitrariness of post-hoc explanations that have been expressed in works like [1] and [16].

2 OVERVIEW: DIFFERENT TYPES OF EXPLANATIONS

ML explanations come in many shapes and forms, and a specialized terminology to describe different explanations has sprung up. Here I briefly describe the different types of explanations and state the names I will be using for them.

ML researchers commonly distinguish between *interpretable models* and *black-box* models. Interpretable models have a mathematical and logical structure that easily allow for explanations to be generated directly from that structure. For example, linear models are commonly considered interpretable because it's easy to explain to even non-technical audiences that x, y, or z are the most important inputs contributing to the linear output, whether the inputs contribute negatively or positively, and so on. It's also easy to use visualizations to illustrate the influences of inputs in linear models.

Black-box models are, roughly speaking, those that are too complex to provide for a simple explanation. They are typically taken to include ensemble models like random forests, and any models that use deep neural networks. Due to the difficulty of generating

[1] For the purposes of this paper, I use "objectivity" in strictly this sense. I don't intend this usage to reflect on other uses of the term, the shifting meanings of which have been documented in works like [9].

FAT '20, January 27–30, 2020, Barcelona, Spain*

ACM ISBN 978-1-4503-6936-7/20/02…$15.00
https://doi.org/10.1145/3351095.3372836

an explanation directly from a black-box model, researchers often resort to generating more interpretable *surrogate models* that mimic a black-box model's behavior, then extracting explanations from the surrogate models. Explanations generated from surrogate models are often called *post-hoc explanations*.

Local explanations are a common type of explanation used to interpret black box models. They focus on explaining, for a particular output, how it can be derived from the input data. They do this without reference to the original model that produced the output, by essentially constructing a surrogate model around the particular input point that produces the expected output.

Global explanations, in contrast, explain how the whole range of outputs is generated from the whole range of inputs. Interpretable models typically provide global explanations. It is also possible to construct surrogate models for black-box models that are global explanations.

Counterfactual explanations have been proposed as a particularly useful form of explanation for communicating ML-generated decisions to impacted users, because of their resemblance to everyday explanations in human conversation [30]. Counterfactual explanations state what would have happened had the input variables been changed in certain ways. These explanations are particularly useful when you want to help the user understand how they can change inputs under their control in order to achieve a different outcome.

3 WHAT WE'RE EXPLAINING

Before discussing desiderata for explanations, it's helpful to first clarify what it is that ML explanations aim to explain. The lack of consensus about desiderata stems partly from a lack of clarity on what the appropriate *explanandum* is—what phenomena we're trying to explain. For example, a local explanation has a different explanandum from a global explanation, because the latter includes the entire range of inputs and outputs for the model in its explanandum, whereas a local explanation focuses on only a single input-output mapping.

I focus on another difference in explananda that has not been explicitly discussed: the extent to which model-world relations are included in the explananda. To understand how the world enters into explanations, consider Figure 1, which depicts relationships between the world, the data input into the model, the model and the model's output.

I constructed this diagram to show that there are many potential explananda that we could pick out from among the diagram's components. The following are an incomplete list:

(1) For a given input, how outputs are generated in the local neighborhood of that input. This is the kind of explanandum targeted by a local explanation. In Figure 1, these explanations have only the *data input* → *model output* relationship as the explananda, and each local explanation covers only a specific data input point.
(2) How the model transforms inputs to outputs, in a *global* sense. Global explanations of interpretable models have this type of explanandum. In Figure 1, these explanations cover the *data input* → *model* → *model output* relationship.
(3) Given a particular input, why it leads to a particular outcome, and how that input can be changed if you want other outcomes. Counterfactual explanations have this type of explanandum. Counterfactual explanations do not include the model's mechanism in their explananda.
(4) How the model captures real patterns in the data. Here, I'm using the term "real patterns" in the abstract, general sense captured in [10].[2] Explanations containing this type of explanandum may relate abstract representations in the data to patterns in the world. For example, a robust interaction between two variables, if it occurs in multiple well-performing models, may shed light on how the variables may be causally related—even if the variables do not have a direct causal relationship, they may share a common cause or act as proxies for other properties that are directly causally connected. For these kind of explanations, the explananda include the whole of Figure 1, because we do want to explain how the flow of *data input* → *model* → *model output* provides insight on the *world* → *model* relationship.

Explanandum 4 is implicit in some types of explanations, but is rarely explicitly articulated in the explainable AI literature. ML models are used in various sciences not just to predict outcomes, but also to help reveal underlying relationships in the world (using data collected from the world as a proxy). In some cases, explanations are evaluated not just based on how they explain the inner workings of the model, but also the extent to which the model provides insight on real relationships in the world.

Using prior philosophical work on how robustness is an indicator of reality, I argue that if we're interested in explanandum 4, then we need to ensure that our explanations are robust. Section 6 provides further epistemic and ethical reasons to prefer explanations that reflect real-world patterns.

4 ROBUSTNESS AS REALITY

Before I go on to survey the robustness of different types of explanations, it's worth taking a step back to ask why we care about robustness and how it may be an indicator that our explanations are latching on to real patterns.

Robustness as a general desideratum has long been advocated by both natural and social scientists [6, 19]. The idea of robustness is indicating reality is often first attributed to the biologist Richard Levins [19], and has since been adopted and refined by philosophers of science [31–33].

For the purposes of ML explanations, we can use a slightly modified version of derivational robustness [15]. As originally formulated, derivational robustness was used to describe certain theorems. The theorems were deemed derivationally robust if they could be derived in multiple [at least partially] independent ways.[3] We can easily extend derivational robustness to apply to ML explanations, by defining robust explanations as those that can be generated

[2] "Real patterns" in this sense includes large-scale, emergent regularities in the phenomena. These, rather than microscopic regularities, would be the type of patterns likely to be found by most ML applications.

[3] The question of what count as independent methods of derivation is complex. See [29] for a discussion. For future work, it would be interesting to explore in what sense the various almost-equally-well-performing models found as part of the Rashomon Effect, as explained in 5.2, can be considered to be independent or partially independent.

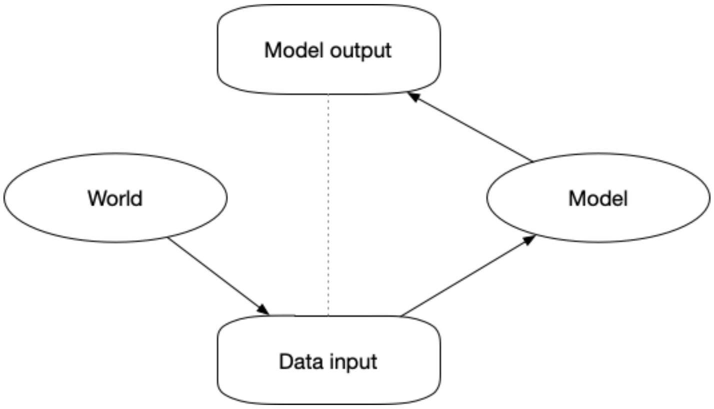

Figure 1: Possible explananda for a model explanation. The dotted line represents the kind of proxy relationship that's produced by post-hoc explanations.

in multiple [at least partially] independent ways, via any of the explanation methods described in Section 2.[4]

Philosophers have offered the following reasons to take robustness as at least an epistemic criterion for when to consider something to be real—that is, a criterion that tells us when we have good evidence that something is real [32]:

- If each model uses different assumptions to derive the same result, that means the result is independent of the differences between those assumptions. This is important if we have reasons to doubt the truth of those assumptions.
- Guarding against errors in the data: If you get the same result by varying some input variable slightly, this means that the result still holds if your the error associated with that input point (whether it comes from data collection or from data processing) is within that slight variation.
- When a result is arrived at by only one method, then the chances of error in each step of that method contribute additively to the overall chance of error for the method: the method is only as reliable as its weakest link. When multiple partially independent methods arrive at the same result, the probability that the result is unreliable is calculated by multiplying the probabilities of each of the methods being unreliable. To put it intuitively, with multiple indepedent methods, all the methods need to be incorrect in order for the result to be unreliable. With only one method, all the steps in the method have to be correct for the result to be reliable. This means that even multiple slightly unreliable methods can lead to a fairly reliable result, whereas one moderately reliable method may not.

We can extend these reasons for robustness to ML explanations by thinking of an explanation as a "result" derived from a modeling process.

I'll next provide some examples of non-robust explanations, as defined by derivational robustness, and explain how, by the reasoning above, they fail to provide good evidence that their content is grounded in real patterns.

5 NON-ROBUST EXPLANATIONS

This section describes some situations where ML explanations lack robustness. In the previous section, I defined robustness in terms of multiple methods arriving at the same "result". Since I'm concerned about *explanations* in this paper, it's worth clarifying that the lack of robustness I'm about to describe is due to multiple methods not producing the same *explanation*—"result" in this context does not refer to model predictions. Indeed, the examples I'm about to discuss are cases where the predictions of the underlying ML models are pretty robust, but the explanations associated with them are not.

5.1 Non-Robust Local Explanations

Locally Interpretable Model-Agnostic Explanations (LIME) [25] and Shapley values [21] are popular local explanation methods that take a complex black-box model and construct locally interpretable models at specific input points. They fall in the class of local explanations because the constructed interpretable model differs at each input point, and a local explanation constructed for one input point does not necessarily hold at another input point.

Alvarez-Melis and Jaakkola showed that LIME and Shapley values lack robustness for non-linear black-box models, in the following sense [2]. They conducted experiments where they slightly perturbed input values that were entered into the black-box model. They found that the surrogate models and the original black box models produced stable output values in response to the perturbed inputs, but that the perturbations often caused the explanations provided by LIME and Shapley to change drastically. In short, explanations generated by LIME and Shapley values are not robust to small changes in input values, even when the outputs of the surrogate models are.

This lack of robustness is troubling for the following reasons.

[4] Some philosophers have raised doubts over the epistemic utility of derivational robustness, on the grounds that the evidence for even derivationally robust theorems still ultimately comes from theoretical assumptions, not empirical evidence [24]. This is less of a worry for my adaptation of derivational robustness—even though ML explanations are derived from models, those models are built on empirical data.

(1) Data is collected from the world with finite precision, and we should allow for that by ensuring that the uncertainty involved in those measurements does not drastically affect the decisions we make. However, it appears that this uncertainty could drastically affect the explanations we use to explain our results, if we use LIME or Shapley values.
(2) A lack of robustness raises questions about the extent to which the surrogate explanations really capture how the black box works, because we normally expect the original black-box model to process neighboring input points in similar ways, especially if those points lead to the same output.[5] The fact that these so-called local explanations cannot in fact reproduce the "localness" of black-box mechanisms is thus troubling. This is particularly pressing because these local explanations only imitate local input-output mappings—they do not claim to simulate the inner workings of the original black-box model. This means we cannot go back to the black box model to verify which of the candidate local explanations (the one generated by the original input point, or the one generated by the slightly perturbed input point) is more descriptive of the black box's mechanics.
(3) We expect phenomena in the world that are similar to have similar explanations—with the exception of a few systems with "chaotic" dynamics, we normally do not expect that perturbing an input into a scientific model slightly results in qualitatively different behavior. The lack of robustness of these local explanations therefore sows doubt about how far they are reflective of real patterns in the world.

As pointed out in [2], in certain contexts, we may not care how far the explanations are informative about the world, if the only reason we're extracting explanations is for debugging the current model—understanding what factors are important in the mathematics of the current model. However, it's easy to think of contexts where it's important to have an explanation that provides insight into the underlying phenomena being modelled. Section 6 describes some plausible contexts where this kind of insight is important.

5.2 The Rashomon Effect

The Rashomon Effect, also known as the multiplicity of good models, is a phenomenon identified by Leo Breiman where for a given set of data, it's possible to construct many equally well-performing models that may differ greatly in their internal structure (and hence in their attendant explanations) [5]. Some researchers argue that this is a good thing [26], but here I will frame it as a lack of robustness.

In the Rashomon Effect, there are multiple models that perform similarly well—their accuracies are so close to one another that we cannot be sure that the differences aren't due to random factors. Each of those models suggests a different explanation, even though they all have similar input-output mappings. We lack convergence of multiple methods on the same explanation—so we lack robustness with respect to the explanations (even if the input-output mappings are robust).

As a concrete example of this lack of robustness, Jiayun Dong and Cynthia Rudin constructed a Rashomon set—the set of almost-equally-accurate models—for the recidivism data set used in the COMPAS algorithm [11]. They find that models in the Rashomon set differed significantly in the importances assigned to certain variables. In particular, they found that the importance of criminal history is lower when the importance of race is higher, and vice versa. In such a case, taking one model out of the Rashomon set to provide "the explanation" would not be an accurate reflection of the patterns in the data—just because race happens to be an unimportant variable in that one explanation doesn't mean that it is objectively an unimportant variable.

This lack of robustness isn't necessarily *unexpected*—if we expect an explanation to explain only a specific model's inner mechanisms, then it is expected that models with different mechanisms will have different explanations. But, referring back to the discussion in Section 3, if we would like model-world relations to be part of the explanation, then these model-specific explanations are not enough. In the next section, I give some examples of contexts where capturing model-world relations is an appropriate desideratum for an ML explanation.

6 WHEN REAL PATTERNS MATTER

I argue that we sometimes see revealing objective patterns as part of the function of ML models. Indeed, this may be why constraints from the real world are commonly placed on models, independently of whether these constraints improve accuracy. For example, one may constrain a model to be monotonic on the basis of domain knowledge [13]. One may also use domain knowledge to pick out one model from several models that are similarly accurate, if one of them better reflects real-world relationships in its internal mechanisms.

But why should we care if our models are picking out real patterns? In certain contexts, it may be appropriate to not care at all. If we care only about the model's predictions, *as given in the model's current form*, then it may not matter to us that the model's mathematics does not reflect real patterns in the data or the world. In such cases, we may want interpretations solely for accountability to end-users—we may want to be able to provide an explanation of how the model processed data and obtained outputs to those impacted by the model, without requiring that these explanations reflect real-world patterns.[6] However, in many cases we want interpretations because we want to better understand the model-world relationship. In other cases where we're interested in improving predictions in a *future* version of the model, we may want to better understand the model-world relationship as a means to the end of making a superior model—so having "purely predictive" intentions does not necessarily absolve you from ensuring that your model is grounded in reality. Indeed, real-world data science is a highly iterative and experimental process where one improves a model based

[5] Indeed, one feature of models using decision tree ensembles (which are generally considered to be black-box models) is that they tend to have smoother input-output relationships than their component decision trees do. In this sense, the smoothness of the input-output relation is related to how black-box they are—ensembling makes the model less interpretable and increases smoothness at the same time.

[6] Note, however, that this depends on how you define "accountability". In some contexts, "accountability" might require that the model is based on our understanding of the real world. In that case, ensuring that your explanation is grounded in real patterns may be necessary for accountability.

on a flawed earlier version of the model, and compatibility with reality is one heuristic that can narrow the space for experimentation and guide one to a better model.[7]

6.1 Hypothesis Generation

Scientific applications are one area where ML explanations are often used to gain insight into real patters in the world. For example, interpretable models that act as surrogates to black-box models can be used to identify patterns that can be combined with physicians' causal models to suggest new hypotheses to investigate [4, 7]. This can be the case even if the models themselves represent mostly non-causal relationships—non-causal information can inform work on causal models.

The process of hypothesis generation can also be important as part of the iterative modeling process itself. If we have robust explanations that we are more confident are not an artifact of arbitrarily selecting one model out of hundreds that perform almost as well, we'll have more confidence going forward that the hypotheses suggested by those explanations are worth paying attention to. In other words, having a more robust explanation narrows the search space of possible model improvements more, as compared to having a less robust explanation.

6.2 Non-Misleading Explanations

Having explanations that pick up on real patterns may also be important if we want to avoid giving users a one-sided picture of what contributed to the decisions that impact them. Suppose we follow Rudin's preferred strategy of picking an interpretable model from the multiplicity of good models, using that to make decisions that affect stakeholders, and providing stakeholders with the explanation associated with that model [26]. While the explanation would certainly be accurate of the model that was used, this strategy still leaves the stakeholder unaware of the fact that there were many other equally good models that would likely have produced the same result but that would produce significantly different explanations for that result. This is not ideal, because it means the stakeholder is unaware of alternative equally-good explanations for the outcome. As I elaborate in Section 7, this can have more severe consequences if "fairwashing" occurs as a result.

This line of argument may conflict with your intuitions somewhat: If we *are* indeed using an interpretable model, and if we provide an explanation to the stakeholder that accurately describes the model's mechanisms, isn't that enough from a transparency perspective? Does the stakeholder really need to know about models that were rejected in the model development process, if those models aren't used at all?

My answer to this is somewhat nuanced: I don't want to reject the idea that sometimes, providing an explanation for only the model that you used, even if there were many others you could have used but didn't, is appropriate given certain applications and certain intended audiences for that explanation. For example, if your intended audience was composed of seasoned data scientists, they might understand that this is a normal part of the model development process, and they might find a full disclosure of the details of that process to be excessive. However, if the audience consists of non-technical people who are looking for an explanation for a decision that affects their lives (for example), providing a unitary "explanation" based on the one model that happened to be picked out of a hundred equally well-performing ones would be, I argue, a form of miseducation.

Explanations ought to be provided with some awareness of how they will be interpreted by their audience, given the social context in which they occur. As Mittelstadt et al. point out, providing an explanation of an ML-generated decision to the user affected by the decision can act as a kind of argumentative support for or justification of that decision [23]. This is particularly likely to be the case when the explanation is for why the user's request for something desirable (e.g. a credit line increase) was denied. Once we realize that explanations in certain contexts are likely to be received as justifications, simply describing the mathematical mechanisms of one particular model begins to seem less adequate. I do not have room in this paper to give a complete analysis of what might form an adequate explanation given its potential uptake as a justification, but I can point to some aspects of non-robust explanations that are unsatisfactory from the point of view of acting as possible justifications:

- We generally expect justifications to be based on real patterns in the world (e.g. the pattern that people with a history of repaying loans on time in the past have social and economic circumstances that make them more likely to repay loans in the future), but as I've argued, the non-robustness of certain explanations makes it unclear that their content reflects real patterns.
- Discursively, if I ask for an explanation of a phenomenon, and just one explanation is provided, with no indication that there are perhaps 100 other explanations that would have served just as well, I would feel that crucial information was being withheld from me. Plausibly, the Rashomon Effect occurs when many features interact with one another in complicated ways, so that some models are able to imitate the contribution of a feature or a combination of features by another seemingly distinct feature or combination of features. The different "explanations" provided are thus different views on the same interactional phenomenon. Providing a "summary" of these different views, for example through something like variable importance curves [11] or model class reliance [12], can provide a fuller picture of what real-world phenomena underlie the prediction.

6.3 Model Debugging and Improvement

Having explanations that reflect real patterns is also better for model debugging and improvement. This may be somewhat contrary to the remark in [2] that non-robust explanations suffice for model debugging. The difference really lies in what we consider within the scope of "debugging".

If an explanation accurately describes the inner mechanisms of a model but there is no evidence that those mechanisms reflect real patterns, then the explanation would be useful for checking that the model behaves as expected according to our understanding of its math. In other words, we can use the explanation to debug

[7] Models of the data science lifecycle acknowledge the iterative nature of the process. See for example [22] and [28].

the model by ensuring that the outputs it produces are consistent with our understanding of its mathematics. However, if we have no evidence that the explanation actually captures real patterns, then we wouldn't expect the explanation to guide us in engineering the model to take advantage of real patterns to improve its accuracy.

In contrast, if we have evidence that our explanation captures objective patterns in the world, we can more confidently use those patterns to debug or improve the model in the following ways:

- Improve feature engineering, using the previous iteration of the model as a guide to which features are important, how to best process features before using them as model inputs, how features interact, and so on.
- Form hypotheses about what other kinds of data would be useful to collect.
- Change the way features are combined in the model.
- Confirm that the model's outputs are consistent with the real-world pattern.

This is an incomplete list. The key point is that learning real patterns gives us insight into the model-world relationship, which means we can learn how to manipulate the variables and relationships in those patterns to use real-world patterns in an informative manner.

7 MORAL HAZARD

Beside concerns about objectivity, the Rashomon Effect also provides a moral hazard in the following way: Knowing that some explanations are more acceptable to end-users than others, organizations may decide to select the model that provides the most acceptable explanation to end-users.[8] For example, they may select the model that appears to assign a very low importance to sensitive factors like race and gender, even if there are many other candidate models that perform similarly to the selected model and that have accompanying explanations that make race and gender seem important. As the original paper on the multiplicity of good models points out, it is common that decision trees with very different structures will have similar accuracies on the same dataset [5]. A significant difference in structure corresponds to a significant difference in the explanation.

A similar moral hazard was exposed in the case of post-hoc explanations. In [1], researchers were able to generate multiple rule lists that approximate an unfair black-box model, then select the subset of rule lists that appear to be more fair. One rule list from this subset would then serve as *the* post-hoc explanation presented to users.

In [16], Lakkaraju and Bastani go one step further in demonstrating both that fairwashing is possible *and* that fairwashing succeeds in improving perceptions. The authors use a black-box model that predicts whether to release defendants on bail. They discover from surveys that their experimental subjects considered race and gender to be inappropriate factors to consider for such decisions. Based on this, they use the MUSE post-hoc explanation generation method [17] to selectively generate two categories of explanations:

- Explanations that cite factors perceived as inappropriate for the prediction purpose, in addition to other factors that are considered appropriate.
- Explanations that cite only factors that are perceived to be appropriate factors to consider in the decision.
- Explanations that cite only factors that are perceived to be neutral.

The authors find that fairwashing works: Study participants found explanations in the second and third groups much more acceptable than explanations in the first group, even though all explanations had been generated for the same black-box model and had similar levels of fidelity to the black-box model's predictions.

The communicative difficulties here are similar to those described in 6.2. A seasoned data scientist might be able to appreciate that a model that claims not to use race as an input variable (for example) may nonetheless be indirectly making use of information about race in some way, by way of a complicated combination of seemingly non-race-related input variables that, together, act as a proxy for race. Thus, communicating to this data scientist that this is the (apparently race-neutral) explanation for a particular decision may not be misleading, because the audience in this case has enough context to understand the limitations of this explanation. However, when the same apparently race-neutral explantion is presented to an audience that is less data science-savvy as *the* explanation for a set of decisions, they are less likely to realize the epistemic limitations of the explanation.[9]

Rudin views the multiplicity of models, and hence of explanations, as a boon because it allows a domain expert to add more constraints to a model without losing accuracy [26]. This is indeed an advantage if there is a domain expert available to do this, and if we can trust the domain expert to add constraints that are based on reality, rather than constraints that are based on the kinds of explanations that are most convenient for the organization. However, this advantage can quickly turn into a disadvantage, from a social point of view, if organizations are incentivized to choose only explanations that "look good" to end-users or regulators, leaving out other explanations that may be just as accurate from the point of view of objective constraints, but that would cause reputational damage. Often, if there are many input variables that are causally intertwined, it's unlikely that domain experts can impose enough reality-based constraints to determine which among several models with significantly different feature importance rankings (for example) is closest to reality. At some point, the organization has to choose among the explanations that the domain expert has deemed consistent with reality, and there is a moral hazard in allowing such choice. As part of the push to consider ML systems as sociotechnical systems rather than as abstract entities in isolation [8, 27], we should consider how social incentives are likely to work when organizations are given the opportunity to make public communications that are "technically" true but that omit context which would enable their audiences to make inferences that, perhaps, work against the organization's interests.

[8] This danger has already been pointed out as a risk for post-hoc explanations, for example in [20] and [23].

[9] Presenting multiple explanations to non-expert end-users may not help much with trust, either. Lakkaraju and Bastani constructed a tool to allow study participants to explore different post-hoc explanations, but found that users largely did not trust the tool [16].

The points just made reflect a broader issue of how providing ML explanations is not sufficient for accountability and transparency. Explanations that accurately reflect a model's mechanisms may be arrived at through a design process (e.g. feature selection, feature engineering, or eliminating other candidate interpretable models in the "Rashomon set" of almost-equally-accurate models) that people impacted by the model's decisions do not understand. Without understanding the design process, it is unclear why they should trust the explanation that comes out of such a design process. Accountability and transparency should apply just as much to the model development process as it does to the abstract model and the data.

8 DOES IT MATTER THAT ML MODELS ARE (USUALLY) NOT CAUSAL?

I've drawn on arguments from the methodology of scientific modeling to argue for the importance of robustness, relying on arguments that robustness is an indicator of reality. One possible objection is that ML models are not like scientific models, because the latter usually try to capture causal relationships, while many ML models are "purely predictive and not causal" [26]. As argued in Section 6.3, though, even for a purely predictive model, understanding how features are related in reality helps in improving the model.

In addition, I'd like to question the predictive-causal dichotomy assumed in this type of objection. Many ML models may not accurately describe the causal relationships in the underlying phenomena in a straightforward sense. For example, decision trees are used to model many types of phenomena that do not actually come about through decision tree-like processes. Nevertheless, the models may still provide insight into real relationships, for example by suggesting to us which features interact heavily with respect to their contribution to a certain outcome. The exact dynamics of the interaction may not be provided by the ML model, but the ML model may at least reveal the shape of the interaction and generate hypotheses that can be tested through other methods.

This is analogous to how "phenomenological" models in science may provide insight into effective macroscopic degrees of freedom, or provide hints for fruitful directions in which to develop more "fundamental" models that directly represent causal processes [14].

Another possibility is that ML explanations can shed light on non-causal aspects of the world. Analogously, some philosophers of science accept a role for non-causal explanations in science, arguing that these types of explanations may, among other things, take advantage of constraints that the systems face, without relying on a causal chain of events [18]. These explanations are no less grounded in facts about the world; they just don't happen to be capturing causal facts.

9 SUMMARY

The ML research literature teems with techniques for constructing explanations, but rarely states the exact explananda we desire. Without a consensus on the appropriate explanada, it's natural that disagreements will occur over the desiderata for an explanation.

I've argued that in certain application contexts, it is appropriate to include model-world relations as part of the explanandum of an ML explanation. This is helpful not just in scientific contexts where the ultimate aim is to learn about the world, but also in contexts where we want to improve a purely predictive model. If model-world relations are part of the explanandum, then we should also require that explanations are robust in the sense that multiple models with similar predictive outcomes converge on the same explanation. This is desirable for epistemic reasons, but also helps prevent a kind of moral hazard where the most socially acceptable explanation out of many equally plausible ones is cherry-picked.

If model-world relations form part of an ML explanation's explanandum, then the "multiplicity of good models" phenomenon, or Rashomon Effect, becomes a problem. Luckily, there are emerging techniques to find robust explanations in the sense I've defined in this paper [3]. Going forward, it would be good to ground the ever-expanding array of new AI techniques in clear definitions of what counts as an explanation and what makes explanations good.

REFERENCES

[1] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. 2019. Fairwashing: the risk of rationalization. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. 161–170. http://proceedings.mlr.press/v97/aivodji19a.html

[2] David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the Robustness of Interpretability Methods. http://arxiv.org/abs/1606.03490

[3] David Alvarez-Melis and Tommi S Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 7786–7795.

[4] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpreting Blackbox Models via Model Extraction. *arXiv e-prints*, Article arXiv:1705.08504 (May 2017), arXiv:1705.08504 pages. arXiv:cs.LG/1705.08504

[5] Leo Breiman. 2001. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci.* 16, 3 (08 2001), 199–231. https://doi.org/10.1214/ss/1009213726

[6] Donald T. Campbell and Donald W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56, 2 (1959), 81–105. https://doi.org/10.1037/h0046016

[7] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

[8] Kate Crawford and Ryan Calo. 2016. There is a blind spot in AI research. *Nature News* 538, 7625 (2016), 311.

[9] Lorraine J. Daston and Peter Galison. 2007. *Objectivity*. Zone Books.

[10] Daniel C. Dennett. 1991. Real Patterns. *Journal of Philosophy* 88, 1 (1991), 27–51. https://doi.org/10.2307/2027085

[11] Jiayun Dong and Cynthia Rudin. 2019. Variable Importance Clouds: A Way to Explore Variable Importance for the Set of Good Models. *arXiv e-prints*, Article arXiv:1901.03209 (Jan 2019). arXiv:stat.ML/1901.03209

[12] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2018. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *arXiv e-prints*, Article arXiv:1801.01489 (Jan 2018), arXiv:1801.01489 pages. arXiv:stat.ME/1801.01489

[13] Maya Gupta, Andrew Cotter, Jan Pfeifer, Konstantin Voevodski, Kevin Canini, Alexander Mangylov, Wojciech Moczydlowski, and Alexander van Esbroeck. 2016. Monotonic Calibrated Interpolated Look-Up Tables. *Journal of Machine Learning Research* 17, 109 (2016), 1–47. http://jmlr.org/papers/v17/15-243.html

[14] Stephan Hartmann. 1995. Models as a Tool for Theory Construction: Some Strategies of Preliminary Physics. In *Theories and Models in Scientific Processes*, William Herfel, Wladyslaw Krajewski, Ilkka Niiniluoto, and Ryszard Wojcicki (Eds.). Rodopi, 49–67.

[15] Jaakko Kuorikoski, Aki Lehtinen, and Caterina Marchionni. 2010. Economic Modelling as Robustness Analysis. *The British Journal for the Philosophy of Science* 61, 3 (2010), 541–567. http://www.jstor.org/stable/40981302

[16] Himabindu Lakkaraju and Osbert Bastani. 2019. "How do I fool you?": Manipulating User Trust via Misleading Black Box Explanations. *arXiv:1911.06473 [cs]* (Nov. 2019). http://arxiv.org/abs/1911.06473 arXiv: 1911.06473.

[17] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 131–138.

[18] Marc Lange. 2016. *Because without cause: Non-causal explanations in science and mathematics*. Oxford University Press.

[19] Richard Levins. 1966. The strategy of model building in population biology. *American Scientist* 54, 4 (1966), 421–431. http://www.jstor.org/stable/27836590

[20] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability, Vol. abs/1606.03490. arXiv:1606.03490 http://arxiv.org/abs/1606.03490 cite arxiv:1607.02531.

[21] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[22] Microsoft. 2017. What is the Team Data Science Process? https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview

[23] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 279–288. https://doi.org/10.1145/3287560.3287574

[24] Steven Hecht Orzack and Elliott Sober. 1993. A Critical Assessment of Levins's The Strategy of Model Building in Population Biology (1966). *The Quarterly Review of Biology* 68, 4 (1993), 533–546. http://www.jstor.org/stable/3037250

[25] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (Eds.). ACM, 1135–1144. https://doi.org/10.1145/2939672.2939778

[26] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[27] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. ACM, New York, NY, USA, 59–68. https://doi.org/10.1145/3287560.3287598

[28] Colin Shearer. 2000. The CRISP-DM Model: The New Blueprint for Data Mining. *Journal of Data Warehousing* 5, 4 (2000).

[29] Kent W. Staley. 2004. Robust Evidence and Secure Evidence Claims. *Philosophy of Science* 71, 4 (2004), 467–488. https://www.jstor.org/stable/10.1086/423748

[30] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *CoRR* abs/1711.00399 (2017). arXiv:1711.00399 http://arxiv.org/abs/1711.00399

[31] Michael Weisberg. 2006. Robustness Analysis. *Philosophy of Science* 73, 5 (2006), 730–742. https://doi.org/10.1086/518628

[32] William C. Wimsatt. 2012. Robustness, Reliability, and Overdetermination (1981). In *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science*, Léna Soler, Emiliano Trizio, Thomas Nickles, and William Wimsatt (Eds.). Springer Netherlands, Dordrecht, 61–87. https://doi.org/10.1007/978-94-007-2759-5_2

[33] Jim Woodward. 2006. Some varieties of robustness. *Journal of Economic Methodology* 13, 2 (2006), 219–240. https://doi.org/10.1080/13501780600733376