Explainable AI and Conceptualizations of “Explainability”

Morgan Klutzke

Submitted to the faculty of the Cognitive Science Program in partial fulfillment of the requirements for Departmental Honors in the degree of Bachelor of Science in the Cognitive Science Program

Indiana University

April 2021

Accepted by the Cognitive Science Faculty, Indiana University, in partial fulfillment of the requirements for conferral of departmental Honors in the Cognitive Science Program

Eduardo Izquierdo, Ph.D.

Thesis Committee

Ehren Newman, Ph.D.

April 19, 2021

Joshua Brown, Ph.D.

Abstract

Explainable AI (XAI) is a subfield of AI research that seeks to close, or at least narrow, the widening gap between the predictive power of our models and how well we understand them. Around the beginning of the 21st century, the importance of interpretability was downplayed and the field neglected. But as models are increasingly applied in high-stakes decision domains like medicine or criminal justice, the demand for explainability has risen dramatically. In this work I provide an overview of the current state of XAI research, with a focus on the importance of understanding. I also identify what I believe to be XAI's greatest weakness: a lack of cohesion. I discuss how the multifarious notions of what it means for a model to be "explainable" lead to inconsistencies in the research, and conclude by proposing a potential solution.

Table of Contents

Models in artificial intelligence and machine learning are making improvements in performance much faster than improvements in interpretability. This has led to the proliferation of models we do not understand making or contributing to decisions we care deeply about. In response to this trend, research in Explainable AI (XAI) has seen a resurgence of activity. This field aims to increase human understandability of AI models, which is no easy feat. AI models, especially deep neural networks, are notoriously difficult to understand. They are often referred to as “black boxes”: systems where the input is known, the output is known, but the internal processes are unknown. But this characterization is misleading—after all, how could we have built the model in the first place if it was truly a black box? XAI has developed ways of investigating the inner workings of these so-called “black box” models, as well as designing model architectures that have inherent interpretability.

However, while work in Explainable AI is extremely important, and much progress has been made in the past couple of decades, the field suffers from a lack of coherence. Different researchers conceptualize “explainability” in different ways. This is not an issue per se, but without acknowledging these differences and discussing them, the field becomes a breeding ground for inconsistencies and miscommunication. As it stands, the problem of creating “explainable” models is under-defined, and the goals of the field are unclear.

This review aims to show both how important Explainable AI research is, and how the lack of coherence impedes its progress. In Part I, I start by giving an overview of XAI, including the motivations behind the research, the obstacles it is facing, and the three general approaches to overcoming those obstacles. Part II then explores the various ways Explainable AI has been conceptualized, bringing inconsistencies to light while also trying to tease out common themes. Finally, I conclude with my own recommendations for the field.

Part I: An Overview of Explainable AI

First, a note on terminology: “interpretability” and “explainability” tend to be used interchangeably throughout the literature, with most treating them as synonymous (e.g. Molnar (2019)), but also some using only one term but not the other (e.g. Kim et al. (2016)), and some attempting to make a distinction between the two (e.g. Roscher et al. (2020)). Following the lead of Carvalho et al. (2019), I acknowledge the lack of consensus, but for the purposes of this paper will use both interchangeably. I also acknowledge the fact that “interpretable” has become a much more popular term than “explainable” since 2004, when van Lent et al. first coined the term “Explainable AI”. But since this still appears to be the most common way for people to refer to the field as a whole, this is the name I will use.

A. Motivations

Better understanding of the models we use is important in a number of ways. The knowledge itself has inherent value, of course, and can be leveraged to build even better models. Even limited understanding of how a model works can greatly increase its utility by informing the modeler about the best situations to use it in, how best to format the inputs, how best to tune the parameters, etc. Beyond this scope, however, our understanding of these systems also impacts how the technology is integrated into society.

Varshney (2015) distinguishes between two kinds of applications for machine learning algorithms, which he calls “Type A” and “Type B” applications. The AI models used in Type A applications support extremely consequential decisions that typically also have human decision-makers in the loop: this includes algorithms used for loan applications, medical diagnoses, or prison sentencing. Type B applications are less consequential and do not require human oversight, with the algorithms automatically making data-driven predictions and taking actions based on those predictions. Content recommendation, streaming video performance optimization, and speech recognition are all examples of Type B applications.

When AI models are used in Type A applications like driving or healthcare, safety is a major concern. In these high-stakes decision domains where human lives are at stake, the performance of the system cannot simply be “good”; it must be as close to perfect as possible. And it needs to be at this level of performance *before* deployment. When the stakes are this high, it would be irresponsible *not* to attempt to understand your model as best you could. Molnar (2019) provides an excellent example: suppose you are building the AI for a self-driving car, and you discover that, for this model, the most salient feature of a cyclist is the presence of two wheels. This knowledge prompts you to consider some of the possible edge cases way before your system is out the door—e.g. a bicycle with side bags that partially occlude the wheels. Understanding the system generally leads to improvements in the system, so when high-stakes decisions are being made, attempting to understand the model’s mechanisms should be non-optional.

Another problem mitigated by explainability is what Doshi-Velez & Kim (2017) refer to as “mismatched objectives”. In some cases what you want your model to do is not easily quantifiable, so you have to define a proxy objective for the model to optimize instead. This is a particularly common occurrence when using evolutionary algorithms, where the models are judged according to a “fitness function”, meant to emulate the actual goals of the programmer. Several examples of misspecified fitness functions are described by Lehman et al. (2020).

Explainability is also the best defense against so-called “Clever Hans" classifiers. These models may seem to perform with high predictive accuracy, but their decision-making strategy is based on spurious correlations found in the training data, and is therefore not generalizable (Clever Hans was a performing horse that appeared to understand arithmetic, but was actually reading the unconscious body language cues of the person asking the question). For example, it took almost a decade for researchers to find out that their models trained on the PASCAL VOC 2007 image dataset for object classification were relying on an artifact for one of the classes: ironically, horses. Apparently many of the photos of horses had a copyright tag in the lower-left corner, which was consistent enough that a model could use it to achieve high classification accuracy. This artifact was discovered by Lapuschkin et al. (2016) after inspecting saliency maps generated via layer-wise relevance propagation, a method from XAI.

Relatedly, explainability can help us prevent unwanted biases from creeping into the model. In some applications, a model may learn patterns from the training data that we do not want it to use to make decisions. For example, a model trained to predict the lengths of prison sentences may learn that black defendants tend to get longer sentences than white defendants. So, the model will use information about the defendant’s race to generate more accurate predictions. This is a case where most people would be willing to sacrifice some of the predictive accuracy to ensure equitability, and it is unlikely that such a model would actually be deployed without anyone noticing this. However, a more insidious example is described by Goodman & Flaxman (2017): suppose you have a model predicting loan repayment, which you use to decide who is approved to take out a loan. You want to err on the side of caution, so you code the model to be risk averse, i.e. it takes confidence intervals into account when making predictions. The population it is trained on consists of two groups: whites and non-whites, which both have an equal chance to default on the loan. However, whites make up 90% of the population. In this case, all else being equal, the model will compute larger confidence intervals for non-whites than whites, and thus will predict a lower chance for them repaying the loan, despite the fact that the two groups have an equal chance of defaulting. Even worse, if the model continues to learn over time, it will gather more and more instances of the overrepresented group, creating even wider gaps in predictions. This compounding bias can happen even with small initial biases toward one group (Goodman & Flaxman, 2017). Actually implementing this model would be discriminatory and irresponsible. But if the model has a “black box” architecture, the bias can be hard to detect.

In light of concerns about transparency, the European Union recently implemented the General Data Protection Regulation (GDPR), which declares, among other things, that all people should have a “right to explanation” (Parliament and Council of the European Union, 2016). This means that if a model or algorithm was used to make a decision that substantially affected your livelihood (e.g. you are denied a home loan), you have the right to know the reasoning behind such a decision. So far, enforcement of these regulations has generally been lackluster. And despite all of the discussions and concerns surrounding this newly-proclaimed “right to explanation” (e.g. Goodman & Flaxman, 2017), it does not appear that any fines or notices invoking this right have been issued as of this writing.

It is also worth mentioning that there are also some situations where you may not care about your model being interpretable. Doshi-Velez & Kim (2017) point out that something like an ad server (software that chooses which ads to show on a website) does not necessarily need to be interpretable, because mistakes or poor performance are not going to have significant consequences (this is an example of what Varshney (2015) would call a Type B application). In this case explainability may be nice to have, but it is relatively unimportant. In other situations, you may not want the system to be explainable, because knowing how the system works makes it too easily manipulatable. There are ways of confusing models, depending on how much you know about how it works. Images can be altered in such a way as to fool an object recognition model, but with changes imperceptible to a human (Szegedy et al., 2014). Explainability methods may inform either the model creator or an adversary about these potential attacks.

B. Obstacles

As mentioned previously, strictly speaking, there is no such thing as a truly black box model. But that does not mean that understanding how an AI model works is always an easy task. What makes it difficult is a combination of the complexity of the model, and the culture surrounding machine learning applications.

For example, some models are actually quite easy to understand. Take a simple linear regression model. A single independent variable is used to predict a single dependent variable by plotting each pair as a point on a 2-dimensional plane. Next, we find a linear function that best fits this data: a straight line that goes as close as possible to as many of the points as possible. Now we can make predictions about new instances of our independent variable based on the values of the dependent variable, by plugging those values into our linear function.

What makes the linear model described above a “machine learning” model is the step where we find a line of best fit. The process of searching for the best line involves minimizing a cost function, usually the sum of squared errors. This can be done by hand, but it is tedious and time-consuming, so whatever software you are using to create your model (e.g. Excel, SPSS, MatLab, R, etc.) will generally have some built-in functionality to compute this for you. Of course, in doing so it obscures what computations the computer is performing. This creates a disconnect between you and the model, which only widens the more complex the model (and thus the optimization function) becomes.

Modern programmers almost always use libraries or packages (e.g. TensorFlow, Keras, scikit-learn, etc.) to construct and train more complex models like deep neural networks. While some knowledge of machine learning techniques is a prerequisite for getting started, the actual implementation process of the models is greatly simplified, such that very little is needed from the programmer to create a decently-performing model. Once the programmer has a ready-to-use dataset, they generally just need to decide what model architecture to use and tune the parameters (though there are tools for automating that, too). To be clear, these libraries are wonderful resources that help to make machine learning much more accessible and much less frustrating. But they also contribute to the disconnect between model and model creator that drives the need for explainability.

In most cases, machine learning engineers are not incentivized to care about explainability, either. Instead, the focus is on predictive accuracy. This may seem counterintuitive given the previous section, where I discussed how understanding your model can help you specify appropriate fitness functions, avoid “Clever Hans” detectors, etc. But if the overall accuracy is the only performance evaluation metric you use, you can get pretty far without making the most optimal decisions.

These obstacles to explainability occur while creating and training the model, but what about after it is trained? For example, take our simple linear regression model from before. We can look at what linear function it came up with, and learn something about our data based on the values it gave for the slope and intercept. In theory, we can do something similar with more complex models: imagine you have a black-box classifier with a finite input space. You could systematically feed your model every possible input vector and measure its prediction for each one. Now you have all the inputs and all the outputs, but you are probably not much closer to understanding how the model works.

For deep neural networks in particular, it may seem intuitive to try to observe the activations of individual neurons or layers and look for patterns. But the representations learned by neural networks tend to be multiscale and distributed, making any individual prediction dependent on both local and global effects that are difficult to generalize (Samek et al., 2021). Feedforward neural networks are also prone to the “shattered gradients” problem, where the deeper the network is, the more the gradients resemble white noise (Balduzzi et al., 2017). This means that small changes in inputs can lead to large changes in the output, again making it difficult to generalize any explanation that could be made of the individual predictions.

C. Approaches

Broadly speaking, there are three approaches to understanding AI models. One approach is to use models that are *inherently interpretable*. These models are, by design, easy to understand. They may also help with understanding the underlying data. Regression models, for instance, are often used by statisticians with no intention of making predictions, instead examining the learned model parameters to make inferences about relationships between variables.

The other approaches in XAI involve developing *post-hoc explanation methods*. In this case the model has already been trained (or is in the process of being trained), and it is probed in some way to assess how it works. These can be further separated into *model-agnostic* and *model-specific* approaches. Model-agnostic methods make few or no assumptions about how the model works, while model-specific methods can only be applied with a specific kind of model and do not generalize to other architectures.

In the rest of this section, I elaborate on the advantages and disadvantages of using these different approaches, and give a few representative examples of each.

1. Interpretable Models

The obvious advantage of inherently interpretable models is that they are generally easy to understand. But if this really is the case, then why doesn't everyone just always use interpretable models? If the context in which you are building the model requires 100% transparency, then you should use interpretable models exclusively. However, in practical applications this is rarely the case, and most people believe in a trade-off between interpretability and predictive accuracy.

Decision trees (which can also be presented as "rule lists" (Wang & Rudin, 2015)) are probably the most straightforward models for a human to interpret (Freitas, 2014), and tend to be the standard that other interpretable models are held up to. Other common interpretable models include regression models, naive Bayes classifiers, k-nearest neighbors methods, k-means clustering, and autoencoders.

2. Model-agnostic Post-hoc Explanations

Model-agnostic methods treat the model as black box, and as such can work with any kind of model architecture. Ribeiro et al. (2016a) advocate for the use of these methods, focusing heavily on the flexibility that model agnosticism provides. This flexibility can be particularly useful in certain situations: for instance, you may want to make a direct comparison between different kinds of models, or develop software that can work with the explanatory output of any model.

At their simplest, model-agnostic explanations show how the model prediction varies with changes in inputs. For example, a partial dependence plot (PDP) is just a graphical depiction of the effect of one or two input features on the model's output. Individual conditional expectation (ICE) plots and accumulated local effects (ALE) plots are also just variations on the basic concept of plotting input features against model output (Molnar, 2019).

Another model-agnostic method is to create surrogate models—interpretable models that are trained on the inputs and outputs of the "black box" model. A global surrogate model is trained on all of the predictions, and thus gives you an explanation for the model as a whole. The Local Interpretable Model-agnostic Explanations (LIME) method instead creates local surrogate models for the purposes of explaining individual predictions (Ribeiro et al., 2016b).

3. Model-specific Post-hoc Explanations

Finally, model-specific methods seem to be relatively rare compared to the others. They are usually specific to neural network architectures. This makes sense, as neural networks are probably the most popular "black box" models. Many of these methods are also specifically designed for image classifiers, as they can produce visualizations of what the network "sees". These can generally be grouped together as "pixel attribution methods", and their resulting visualizations can all be called "saliency maps" (as is done by Molnar (2019)).

One example is layer-wise relevance propagation (introduced by Bach et al. (2015)), which can identify, given some input vector, which elements in that vector were the most relevant to the predicted class. It does this through a backpropagation-like procedure, where relevance values are fed from the output layer back toward the input layer, but with the constraint that relevance is conserved—each neuron must send to the lower layer as much relevance as it received from the higher layer (see Montavon et al. (2019) for more details). If the input vector is a pixel array, then you can just color the pixels in the image according to their resulting relevance score, and you have a saliency map showing you which parts of the image were most relevant to its classification. Layer-wise relevance propagation is a particularly notable method because even though it was originally implemented with standard deep feed-forward neural networks, the basic idea is generalizable enough that it has been applied in convolutional neural networks and even recurrent networks (Arras et al., 2019).

Another method is the “deep dream” technique, generally applied to convolutional neural networks. This is specific to neural networks performing image classification, but it is not a pixel attribution method. Instead, it effectively inverts the neural network, adjusting an image to maximize the activation of a given neuron (Mahendran & Vedaldi, 2015). This creates a visualization of that neuron’s internal representation, or what the neuron looks for in an image. Mordvintsev et al. (2015) provide an example of using this method to ensure that an image classifier has learned the most important distinguishing features of an object class. They found that a network they had trained on images of dumbbells learned most of the right features, but their model also thought that a key property of a dumbbell was the presence of a muscular arm holding it. This was almost certainly an artifact in the training data, and would have affected the model’s ability to generalize to new images if it was not caught.

Part II: Explaining “Explainability”

What remains unclear at this point is what it really means for a model to be “explainable”, or “interpretable”, or any number of other similar descriptors. A review of the literature suggests that this is not a unified concept, but a benchmark that shifts depending on the researchers’ expectations. For example, consider these two definitions of “interpretability”:

1. “...a method is interpretable if a user can correctly and efficiently predict the method’s results” (Kim et al., 2016).
2. “...systems are interpretable if their operations can be understood by a human, either through introspection or through a produced explanation” (Biran & Cotton, 2017).

Although they both discuss interpretability, these two papers set different standards for what an interpretable model should be. The first definition focuses on *prediction.* In other words, the explanation takes the form of a formula of some sort that can be used to replicate the model’s outputs, though the formula itself may invite confusion. The second definition focuses on *understanding*. In this case the explanation may or may not allow you to accurately predict the outputs of the model, but it does provide a qualitative window into the model’s operations. These two kinds of explanations clearly reflect different facets of understanding, but the field currently does not distinguish between them. Attempts to delineate between such concepts have been earnestly undertaken by several groups (e.g. Lipton (2018), Doran et al. (2017), Roscher et al. (2020), Montavon et al. (2018)), but they tend to conflict with each other, and ultimately none of these frameworks seem to have stuck.

This misalignment across papers appears to reflect varying expectations about what an “explanation” should be, who provides it, and who receives it. In this section we will explore what these different expectations are and how they sometimes come into conflict.

Let’s start by establishing a basic definition, which hopefully will not invite objection: *an explainable model is a model that is able to be explained*. My purpose in explicitly stating this is to provide a common starting point for conceptualizing explainability, and also to invite the reader to consider what natural follow-up questions you would have if someone actually gave this to you as a definition. You would probably want to know:

- What exactly is being explained?
- Who is explaining it?
- Who is receiving the explanation?
- How is it explained?

You may have other questions (e.g. what counts as a model in this context?) but we will focus on these four in this paper. We will look at each question in turn, evaluating the different possible answers, and noting how they change depending on who you ask and what problems they are trying to address.

A. What exactly is being explained?

First of all, some established methods in XAI—generally those classified as model-agnostic—are really not so much explaining the *model* as they are explaining *patterns in the training data* that the model has picked up on (e.g. PDPs, surrogate models, etc). This is still useful information, of course, but it is not the same thing as understanding the model itself. In fact, depending on your conceptualization of what the goal of Explainable AI is, the fact that some methods are "model-agnostic" should seem absurd. How could you be agnostic toward the very thing you are trying to explain?

I will also mention that model-agnostic methods generally rely only on the model's inputs and outputs, which allows them to be applied to any model, but also tells you very little about what computations the model actually performs. This can lead to inaccurate explanations. Rudin (2019) suggests that an example of this is the criticism from criminologists (Flores et al., 2016) levied against ProPublica's analysis of the COMPAS recidivism model (Angwin & Larson, 2016). The ProPublica article accused COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) of depending on race as a factor in its predictions of individuals' future criminal activity. Because COMPAS is a proprietary algorithm, they were forced to treat it as a black box, and their approach was effectively equivalent to the global surrogate method (even if they do not refer to it as such): create an interpretable logistic regression model that approximates the predictions of COMPAS. They concluded that COMPAS exhibits racial bias: black defendants were more likely to be false positives (misclassified as recidivist) and white defendants were more likely to be false negatives (misclassified as non-recidivist). However, COMPAS is not a binary classifier, and imposing a false dichotomy on its predictions both gave biased results, and reflected a misunderstanding in how its predictions are actually used to make decisions (see Flores et al. (2016) for an exhaustive criticism and reanalysis). In this situation, the initial assumptions about the model impacted how it was analyzed and explained, thus invalidating the explanation. This is a hazard affecting all model-agnostic explanations (and probably other explanation methods as well) that should not be ignored.

These problems aside, there are still several different ways to explain a model, in terms of what exactly is being explained. I will outline these in the rest of this section.

First, the explanation may apply to a model *instance* or a model *class*. Borrowing from object-oriented programming, a model *class* defines the general architecture of a type of model (e.g. a decision tree, a neural network, etc). A model *instance* is an actual trained model that belongs to a class, but has specific parameters. So, for instance, if I tell you that the output of a perceptron is just the dot product of the input vector and a vector of weights, plus some bias, compared to a threshold value, that would be an explanation of the perceptron model *class*. But if I tell you that this perceptron arrived at a particular output because it has a larger weight value for this input feature than that input feature—that is an explanation of a perceptron *instance*. Both kinds of explanations are important: it is difficult to understand an instance without understanding its class. However, most methods in XAI focus on explaining instances and assume familiarity with the model classes. We will return to this in the discussion of who receives explanations.

There are also *global* and *local* explanations. A *global* explanation will provide us with information about the model as a whole, e.g. its internal representations. A *local* explanation will instead focus on a single prediction, or a small subset of the possible predictions. Global and local surrogate models are examples of each.

Lastly, there are different *levels* of explanations (inspired by Lipton's (2018) levels of transparency). An *algorithmic* explanation describes the model at the level of its learning algorithm. Proof that a model will converge to an optimal solution would be an algorithmic explanation. The explanation may instead address some *component* of the model—e.g. a single parameter or calculation (or a small subset). The deep dream method produces component explanations of a CNN by focusing on either an individual neuron, or a single layer of neurons. Finally, a *holistic* explanation will characterize the model as a whole. Surrogate models are holistic explanations, since they can produce implications about what calculations a model might be performing, but can only approximate to the level of the entire model.

B. Who is explaining it?

Explanations can be thought of as a social interaction between the explainer and the explainee. But in Explainable AI, it is rarely clear who is taking on these roles. Here we focus on the identity of the explainer.

For any of the existing methods in XAI, there needs to be some kind of intermediary between the method and the explainee. Some may call inherently interpretable models "self-explainable" (e.g. Samek et al., 2021). But even for decision trees, whose operations are about as transparent as any model can get, there may be follow-up questions that cannot be answered just from looking at a diagram (e.g. how did it choose where to split?). Similarly, any post-hoc method will still need some additional explanation, to explain the method itself. There is no such thing as a self-evident explanation—someone or something must take on the role of the "explainer", though how important that role is will depend on the method of explanation.

Nearly always the explainer will need to be a human familiar with the method. The only exception is when the model is embedded into some larger software application which can produce explanations. In this case the explainer is the software itself rather than a human. Examples of this include van Lent et al.'s (2004) explainable AI system for their military training simulation, which the user can interact with during a debriefing phase after each mission, or Swartout et al.'s (1991) advice system for improving users' Lisp code. In both of these applications, the user can ask questions to and receive explanations from the program.

C. Who is receiving the explanation?

The “end user” for explanations can fall into one of three broad categories, depending on the application:

1. The AI community: people who are already familiar with the field. This includes the creator of the model being explained.
2. Decision-makers: people using the model to inform consequential decisions (i.e. Type A applications). For example, a doctor using a model to help make a diagnosis.
3. Laypeople: other people who are not necessarily familiar with AI technology, but whose lives are affected by it.

The lines defining these categories are fuzzy, but they give us a way to identify who is intended to receive the explanations from different XAI methods. Most methods seem to be geared toward the AI community—they rely on some amount of assumed knowledge to be comprehensible, and they rarely show up outside of research papers. They tend to focus on the problems the model creators may care about, like selecting features or handling edge cases. But many of the problems that XAI is supposed to solve involve explaining models to people in the other two categories, who do not have this kind of expertise. This creates a gap between what the field states that they want, and what they actually produce.

To some extent, this is a problem with outreach: not enough people use XAI methods in real-world applications (see “Obstacles” section above). If they did, there may be more pressure for those developing these methods to make their outputs more useful and understandable to a wider group of people. But, at the same time, it would be easier to convince more people to use XAI methods if they were already easily understandable. This leads us to our last question.

D. How is it explained?

Several papers in XAI attempt to devise some objective, axiomatic measures of explanation quality (e.g. Murdoch et al. (2019), Montavon et al., (2018)). While their intentions are good, they rarely take into account the social or cognitive aspects of explanations (though a notable exception is Miller (2019), who provides an extraordinary review of research from the social sciences on explanations, and how those findings might be applied in XAI). Explaining anything is a balancing act between being accurate and being understandable. An explanation that is only concerned with accuracy will end up being just as complex as the model it is trying to explain. An explanation focused solely on understandability will oversimplify to the point that it is no longer a faithful representation of the model. A “good” explanation will lie somewhere in the middle. Where, exactly, in the middle will depend on the context: what the explanation is for and who is receiving it are inseparable from its evaluation.

Additionally, the researchers developing methods in XAI are prone to the curse of knowledge—a cognitive bias that makes it difficult for experts to remember what it was like to be a novice (Camerer et al., 1989). Accordingly, they should not be the ones evaluating their own explanations. Instead, the efficacy of explanation methods should be verified with human subjects, and diverse perspectives from fields like human-computer interaction should be encouraged—especially when the intended recipients are laypeople.

Conclusion

I have already mentioned a few of my suggestions for future work, like exercising caution with model-agnostic methods, considering the social and cognitive aspects of explanations, and inviting perspectives from other fields. However, none of these directly address the core issue at hand: “explainability” is a multifaceted concept. Different people will have different conceptualizations of explainability, and different expectations of what an explainable AI model looks like. The diversity of perspectives is not the problem—in fact, it is an inherent benefit to working collaboratively with other people. But as long as the variety in ideas goes unacknowledged, research in the field will continue to be incohesive.

My proposal, then, is to recognize this diversity by establishing a more specific and comprehensive lexicon that can account for all of the different facets of explainability. To be clear, having precise, exact definitions is neither necessary nor sufficient for scientific progress. For example, take the field of Cognitive Science—we don’t have a standard definition for “cognition”, despite the fact that it is the core object of inquiry that characterizes our field. No, it is *because* we study cognition that there is no standard definition; part of what drives the field is the question of what counts as cognition and what does not. It may be that cognition is not a “natural kind” in the philosophical sense, and debating the boundaries of this arbitrary category will not yield meaningful results. However, the *process* of debating, of attempting to formulate some axiomatic definition of cognition, can still be a fruitful exercise that guides us toward new questions and new ways of thinking.

This is not the case for Explainable AI. What motivates the field is not so much the question of what an explainable model is or should be, but rather solving the problems described above (see “Motivations”). Having standardized definitions here would not be constraining in the way it would for Cognitive Science. Instead it would allow researchers to clearly communicate with each other about what problems they are working on and the approaches they are taking.

As mentioned earlier, several papers have tried to do something like this (e.g. Lipton (2018), Doran et al. (2017), Roscher et al. (2020), Montavon et al. (2018)). But in each case, while their recommendations are thoughtful, their suggested terminology has yet to be widely adopted. They also sometimes conflict with each other, providing further evidence for differences in conceptualizations. Ultimately, if the field is to establish a more consistent vocabulary, I do not think that it will come from the work of just a few authors. If everyone is expected to use the new terminology, everyone needs to have a say in how it is defined.

The best way to hear from as many voices as possible would be to hold some kind of conference with the explicit objective of defining a new lexicon. This event would attract people with the strongest opinions, but also require them to defend those opinions to their peers. As long as the conference was held by a respected organization and attended by notable researchers in the field, even those who did not attend would likely feel some pressure to conform to the new terminology, or at least acknowledge its existence.

Although the stated intention of such a conference would be to establish a common vocabulary, the most important discussions it would provoke would not really be about definitions or terminology. They would be about what expectations and goals we set for ourselves and for our field. Having clearly-defined terms is just a way of making those expectations explicit, so they can more easily be communicated, discussed, and criticized.

References

Angwin, J., & Larson, J. (2016). Machine bias. *ProPublica.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arras, L., Arjona-Medina, J., Widrich, M., Montavon, G., Gillhofer, M., Müller, K., Hochreiter, S., & Samek, W. (2019). Explaining and interpreting LSTMs. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L., & Müller, K. (Eds.), *Explainable AI: Interpreting, explaining, and visualizing deep learning* (pp. 211-238). Springer, Cham. https://doi.org/10.1007/978-3-030-28954-6_11

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE 10*(7). https://doi.org/10.1371/journal.pone.0130140

Balduzzi, D., Frean, M., Leary, L., Lewis, J. P., Ma, K. W., & McWilliams, B. (2017). The shattered gradients problem: If resnets are the answer, then what is the question?. *Proceedings of the 34th International Conference on Machine Learning*, in Precup, D., & Teh, Y. W. (Eds.), *Proceedings of Machine Learning Research, 70,* 342-350. http://proceedings.mlr.press/v70/balduzzi17b.html

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. *IJCAI-17 Workshop on Explainable AI (XAI). 8*(1), 8-13. http://www.cs.columbia.edu/~orb/papers/xai_survey_paper_2017.pdf

Camerer, C., Loewenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy, 97*(5), 1232-1254. https://doi.org/10.1086/261651

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics, 8*(8), 832. https://doi.org/10.3390/electronics8080832

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv.* https://arxiv.org/abs/1710.00794

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. https://arxiv.org/abs/1702.08608v2

Flores, A. W., Bechtel, K., & Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to “Machine bias: There’s software used across the country to predict future criminals. And it’s biased against blacks.” *Federal Probation, 80*(2), 38-46. https://www.uscourts.gov/sites/default/files/80_2_6_0.pdf

Freitas, A. A. (2014). Comprehensible classification models: A position paper. *ACM SIGKDD Explorations Newsletter, 15*(1), 1-10. https://doi.org/10.1145/2594473.2594475

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine, 38*(3), 50-57. https://doi.org/10.1609/aimag.v38i3.2741

Kim, B., Khanna, R., & Koyejo, O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29 (NIPS 2016)* (pp. 2280-2288). https://proceedings.neurips.cc/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf

Lapuschkin, S., Binder, A., Montavon, G., Müller, K., & Samek, W. (2016). Analyzing classifiers: Fisher vectors and deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2912-2920). https://doi.org/10.1109/CVPR.2016.318

Lehman, J., Clune, J., Misevic, D., Adami, C., Altenberg, L., Beaulieu, J., Bentley, P. J., Bernard, S., Beslon, G., Bryson, D. M., Cheney, N., Chrabaszcz, P., Cully, A., Doncieux, S., Dyer, F. C., Ellefsen, K. O., Feldt, R., Fischer, S., Forrest, S.,... Yosinski, J. (2020). The surprising creativity of digital evolution: A collection of anecdotes from the evolutionary computation and artificial life research communities. *Artificial Life, 26*(2), 274-306.

https://doi.org/10.1162/artl_a_00319

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM, 61*(10), 36-43. https://doi.org/10.1145/3233231

Mahendran, A., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (5188-5196).

Miller, T. (2019). Explanations in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Molnar, C. (2019). Interpretable machine learning: A guide for making black box models explainable. https://christophm.github.io/interpretable-ml-book/

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K. (2019). Layer-wise relevance propagation: An overview. In Samek, W., Montavon, G., Vedaldi, A., Hansen, L., & Müller, K. (Eds.), *Explainable AI: Interpreting, explaining, and visualizing deep learning* (pp. 193-209). Springer, Cham. https://doi.org/10.1007/978-3-030-28954-6_10

Montavon, G., Samker, W., & Müller, K. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing, 73*, 1-15. https://doi.org/10.1016/j.dsp.2017.10.011

Mordvintsev, A., Olah, C., & Tyka, M. (2015, June 17). *Inceptionism: Going deeper into neural networks.* Google AI Blog. https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences, 116*(44), 22071-22080. https://doi.org/10.1073/pnas.1900654116

Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union,* 1-88. http://data.europa.eu/eli/reg/2016/679/oj

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-agnostic interpretability of machine learning. *arXiv*. https://arxiv.org/abs/1606.05386

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). “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* (pp. 1135-1144). https://doi.org/10.1145/2939672.2939778

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access, 8*, 42200-42216. https://doi.org/10.1109/ACCESS.2020.2976199

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

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE, 109*(3), 247-278. https://doi.org/10.1109/JPROC.2021.3060483

Swartout, W., Paris, C., & Moore, J. (1991). Explanations in knowledge systems: design for explainable expert systems. *IEEE Expert, 6*(3), 58-64. https://doi.org/10.1109/64.87686

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *arXiv.* https://arxiv.org/abs/1312.6199v4

Van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. In *Proceedings of the National Conference on Artificial Intelligence* (pp. 900-907). https://www.aaai.org/Papers/IAAI/2004/IAAI04-019.pdf

Varshney, K. R. (2015). Data science of the people, for the people, by the people: A viewpoint on an emerging dichotomy. In *Proceedings of the Bloomberg Data for Good Exchange*

Conference. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.1063.3761&rep=rep1&type=pdf

Wang, F., & Rudin, C. (2015). Falling Rule Lists. *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, in Lebanon, G., & Vishwanathan, S. V. N. (Eds.), *Proceedings of Machine Learning Research, 38*, 1013-1022. http://proceedings.mlr.press/v38/wang15a.pdf