Explaining Explanations in AI

Brent Mittelstadt
brent.mittelstadt@oii.ox.ac.uk
University of Oxford
The Alan Turing Institute

Chris Russell
crussell@turing.ac.uk
University of Surrey
The Alan Turing Institute

Sandra Wachter
sandra.wachter@oii.ox.ac.uk
University of Oxford
The Alan Turing Institute

ABSTRACT

Recent work on interpretability in machine learning and AI has focused on the building of simplified models that approximate the true criteria used to make decisions. These models are a useful pedagogical device for teaching trained professionals how to predict what decisions will be made by the complex system, and most importantly how the system might break. However, when considering any such model it's important to remember Box's maxim that "All models are wrong but some are useful." We focus on the distinction between these models and explanations in philosophy and sociology. These models can be understood as a "do it yourself kit" for explanations, allowing a practitioner to directly answer "what if questions" or generate contrastive explanations without external assistance. Although a valuable ability, giving these models as explanations appears more difficult than necessary, and other forms of explanation may not have the same trade-offs. We contrast the different schools of thought on what makes an explanation, and suggest that machine learning might benefit from viewing the problem more broadly.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Cognitive science**; **Machine learning**; • **Human-centered computing** → *HCI theory, concepts and models*;

KEYWORDS

Interpretability; Explanations; Accountability; Philosophy of Science

ACM Reference Format:
Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining Explanations in AI . In *FAT* '19: Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29–31, 2019, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3287560.3287574

1 INTRODUCTION

As we deploy automated decision-making systems in the real world, questions of accountability become increasingly important.[1] For system builders questions such as "Is the system working as intended?", "Do the decisions being made seem sensible?" or "Are we conforming to equality regulation and legislation?" are important, while a subject of the decision-making algorithm may be more concerned with topics such as "Am I being treated fairly?" or "What could I do differently to get a favourable outcome next time?"

These issues are not unique to computerised decision-making systems, but with the growth of machine learning based systems they have become even more important (Bodo et al., 2017; Kroll et al., 2016; Nissenbaum, 1996; Olhede and Wolfe, 2018; Pasquale, 2015; Selbst and Barocas, 2018; Veale and Edwards, 2018). What distinguishes machine learning is its use of arbitrary black-box functions to make decisions. These black-box functions may be extremely complex and have an internal state composed of millions of interdependent values. As such, the functions used to make decisions may well be too complex for humans to comprehend; and it may not be possible to completely understand the full decision-making criteria or rationale.

Under these constraints, it becomes an open question as to what forms of explanation are even possible that can answer the earlier questions. As such, one of the most striking aspects of research into explainable AI (xAI) is how many different people, be they lawyers, regulators, machine learning specialists, philosophers, or futurologists, are all prepared to agree on the importance of explainable AI. However, very few stop to check what they are agreeing to, and to find out what explainable AI means to other people involved in the discussion (Lipton, 2016).

This gap of expectations is largest between machine learning, which has essentially re-purposed the term "explanation," and the fields of law, cognitive science, philosophy and the social sciences (which we refer to here collectively as the 'explanation sciences'), where it has a relatively well-defined technical meaning and a plethora of research on types of explanations, their purposes, and their social and cognitive function.[2] Work on xAI currently occupies only one or two small branches of this diverse research landscape. Specifically, the vast majority of work in xAI produces simplified approximations of complex decision-making functions. We argue that these approximations function more like scientific models than the types of scientific and 'everyday' explanations considered in philosophy, cognitive science, and psychology.

In this paper we examine the extent of this gap between xAI and the 'explanation sciences'. We do so by first reviewing methods for producing explanations in xAI, and explain how they are generally more akin to scientific modelling than explanation giving. If this comparison holds, it follows that the majority of currently available methods will, at best, produce locally reliable but globally

[1]This work was supported by The Alan Turing Institute, EPSRC grant EP/N510129/1, and the British Academy, grant PF170151.

FAT'19, January 2019, Atlanta, Georgia USA*

ACM ISBN 978-1-4503-6125-5/19/01…$15.00
https://doi.org/10.1145/3287560.3287574

[2]The dichotomy between explanations in machine learning and elsewhere is well illustrated by the recent work by Miller et al. (2017) who found that none of the papers on a curated reading list for xAI made use of work from the social sciences on explanations.

misleading explanations of model functionality. We then examine research on explanations in philosophy, cognitive science, and the social sciences which suggests that 'why-questions' (e.g. 'why did the model exhibit that behaviour?') require explanations that are contrastive, selective, and socially interactive. On this basis, we argue that, if xAI is to produce methods that make algorithmic decision-making systems more trustworthy and accountable, the field's attention must shift to the development of interactive methods for post-hoc interpretability that make it easier to contest algorithmic decisions, and facilitate informed dialogue between users, developers, algorithmic systems, and other stakeholders.

2 A BRIEF PRIMER ON EXPLANATIONS IN PHILOSOPHY

Our interest in the philosophical treatment of explanations is based on observation of the development of the field of xAI, or research addressing interpretability and explainability in machine learning. Our aim is to determine whether xAI is heading in a direction in which explanations can be produced that allow affected parties, regulators, and other non-insiders to understand, discuss, and potentially contest decisions made by black-box algorithmic models. So our primary question is: which types of explanations are currently being produced by xAI? And, are these explanations actually useful to the individuals (or proxies thereof) affected by black-box decisions?

Explanations, and more broadly epistemology, causality, and justification, have been the focus of philosophy for millennia, making a complete overview of the field unfeasible. What follows is a brief review of key distinctions and terminology relevant to our interest in methods for providing explanations of black-box algorithmic models and decisions. This primer broadly follows the structure established in a recent review of the contribution of the social sciences to xAI (see: Miller (2017)).

Briefly, the xAI community investigates interpretability (or explainability) and ways of providing explanations of algorithmic models and decisions. 'Interpretability' refers to the degree of human comprehensibility of a given 'black-box' model or decision (Lisboa, 2013; Miller, 2017). Poorly interpretable models "are opaque in the sense that if one is a recipient of the output of the algorithm (the classification decision), rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs" (Burrell, 2016, p.1).

In contrast, 'explanation' refers to numerous ways of exchanging information about a phenomenon, in this case the functionality of a model or the rationale and criteria for a decision, to different stakeholders (Lipton, 2016; Miller, 2017). Explanations of machine learning models and predictions can serve many functions and audiences. Explanations can be necessary to comply with relevant legislation (Doshi-Velez et al., 2017), verify and improve the functionality of a system (i.e. as a type of 'debugging'; (Kulesza et al., 2015)) and help developers and humans working with a system learn from it (Samek et al., 2017), and enhance the trust between individuals subject to a decision and the system itself (Citron and Pasquale, 2014; Hildebrandt and Koops, 2010; Zarsky, 2013). As these purposes suggest, explanations can be offered to expert developers, professionals working in tandem with a system (e.g. expert labelers of training cases; see Berendt and Preibusch (2017) for an overview of human actors involved in algorithmic decision-making), and to individuals or groups affected by a system's outputs (Mantelero, 2016; Weller, 2017).

Our interest here is solely in ways of providing explanations. Returning to philosophy, types of explanations can be distinguished according to their completeness, or the degree to which the entire causal chain and necessity of an event can be explained (Ruben, 2004). Often this is expressed as the difference between 'scientific' and 'everyday' explanations (both of which deal with causes of an event; e.g. Miller (2017), or 'scientific' (full) and 'ordinary' (partial) causal explanations (Ruben, 2004). Miller argues that 'everyday explanations' address "why particular facts (events, properties, decisions, etc.) occurred," rather than general scientific relationships (Miller, 2017, p. 5).

While these distinctions hide significant nuance, they matter insofar as they constrain the scope of our focus in discussing explanations in AI. In recent calls for explanations in AI, and in work on interpretability in machine learning more broadly, explanations are requested in connection to a particular entity, be it a specific decision, event, trained model, or application. The explanations requested are thus not full scientific explanations, as they need not appeal to general relationships or scientific laws, but rather at most to causal relationships between the set of variables in a given model (Woodward, 1997). As such, xAI is effectively calling for everyday explanations either of how a trained model functions in general, or how it behaved in a particular case.

3 EXPLAINABLE AI

Much recent work has been dedicated to rendering machine learning models interpretable or explainable. Two broad aims of work on interpretability have been recognised in the literature: transparency and post-hoc interpretation. Transparency addresses how a model functions internally, whereas post-hoc interpretations concern how the model behaves (Lepri et al., 2017; Lipton, 2016; Montavon et al., 2017). Transparency can be further specified according to its target. Respectively, a mechanistic understanding of the functioning of the model (simulatability), individual components (decomposability), and the training algorithm (algorithmic transparency) can be sought. Models can be rendered transparent by explanations at a minimum of three levels: "at the level of the entire model, at the level of individual components (e.g., parameters), and at the level of a particular training algorithm" (Lepri et al., 2017). A model, its component parts, or its training/learning algorithm can thus be said to be transparent if their functionality can be comprehended in their entirety by a person (Lipton, 2016).

Post-hoc human interpretable explanations of models and specific decisions do not seek to reveal how a model functions, but rather how it behaved, and why. According to Lipton (2016), approaches to post-hoc interpretability include verbal (natural language) explanations (e.g. McAuley and Leskovec (2013)), visualisations and interactive interfaces (e.g. Simonyan et al. (2013); Tamagnini et al. (2017)), local explanations or approximations (e.g. Fong and Vedaldi (2017); Ribeiro et al. (2016)), and case-based explanations (e.g. Caruana et al. (1999); Kim et al. (2014)).

Natural language explanations can consist of "textual or visual artefacts that provide qualitative understanding of the relationship" between features of an input (e.g. words in a document) and the model's output (e.g. a classification or prediction; (Ribeiro et al., 2016). Visualisation techniques can visually demonstrate the relative influence of features or particular pixels (e.g. in the case of an image classifier), or provide an interface for users to explore textual or visual explanations (Poulin et al., 2006; Tamagnini et al., 2017). Local explanations seek to explain how a fixed model leads to a particular prediction, either by fitting a simpler, local model around a particular decision (Ribeiro et al., 2016), or by perturbing variables to measure how the prediction changes (Adler et al., 2016; Datta et al., 2016; Simonyan et al., 2013). Case-based explanation methods (Caruana et al., 1999; Kim et al., 2014) for non-cased based machine learning involve using the trained model as a distance metric to determine which cases in the training data set are most similar to the case or decision to be explained. These training cases can then be shared with parties affected by the decision.

Despite this variety of approaches, a significant amount of the xAI community now pursues methods to retro-fit local or approximate models over more complex algorithms. These simplified models approximate the true criteria used to make decisions (e.g. Baehrens et al. (2010); Ribeiro et al. (2016); Selvaraju et al. (2016); Simonyan et al. (2013)). Broadly speaking there are two widely used classes of model *(i)* Linear or Gradient-based approximations that assign a single importance weight to each feature (be it someone's age, or a particular pixel in an image) and *(ii)* Decision tree-based methods that use nested sets of yes/no decisions to approximate classifiers.

These methods can both be applied to create approximations at a global[3] or local[4] level. Historically, much work has focused on global approximations of models (e.g. Craven and Shavlik (1996); Martens et al. (2007); Sanchez et al. (2015), including approaches based on clustering (Chen et al., 2016), integer programming (Zeng et al., 2017), and rule lists (Wang and Rudin, 2015)). In contrast, local approximations are accurate representations only of a specific domain or 'slice' of a model. A trade-off inherently occurs between the insightfulness of the approximated model, the simplicity of the presented function, and the size of the domain to which is applies and remains valid (Bastani et al., 2017; Lakkaraju et al., 2017).

No matter the approach taken in xAI, reflexivity is needed to ensure the community actually works towards its normative and practical goals to render models holistically transparent or provide high-quality post-hoc interpretations of model behaviour. Critical questions must be repeatedly asked and answered. For example, will the methods developed make machine learning models more interpretable? More trustworthy to users? More accountable? And to whom will explanations be accessible, comprehensible, and useful?

Answering such questions requires considering the methods developed in xAI in the context of prior work in fields addressing such normative and social questions. Local and approximation models may in fact resemble existing, well-known approaches to explanations in the 'explanation sciences', which would provide insight into their practical value and limitations for users, developers, and other stakeholders going forward.

[3]Offering a simplified model or a set of simplified models that approximates decisions made for all possible datapoints.

[4]A simplified model that only approximates decisions made about a few datapoints, typically only a single exemplar.

3.1 Scientific Modelling and Explainable AI

We believe that the closest analogue for the bulk of methods currently occupying xAI researchers lies in the use of scientific modelling, or the building of approximate models that are not intended to capture the full behaviour of physical systems but rather to provide coarse approximations of how the systems behave. These approximations are useful to experts both for pedagogical purposes and for making reliable predictions of how the system might behave over a restricted domain, but can be misleading when presented as an explanation of how the model functions to a lay user.

One famous example of this problem is Newtonian physics, which is taught first to schoolchildren and provides a good enough description for much day-to-day engineering, but that famously breaks down as an approximation when high-precision is required at either very large or very small scales, where either general relativity or quantum physics are necessary.[5] Both general relativity and quantum physics are also examples of such models. Although extremely accurate in their domains, outside of them they break down, and a unified model of physics that is accurate at all scales is still being sought.

Much scientific theory can be understood as the use and characterisation of such models (Box, 1979; Frigg, 2006; Herfel et al., 1995). Although any physical system can be understood in terms of the emergent properties of subatomic particles, such descriptions are neither human comprehensible nor computationally feasible. Instead, scientists deal in local approximations that provide accurate descriptions of the phenomena they are interested in, but which may prove inaccurate in a larger domain.

It is in this context that Box's maxim, "All models are wrong, but some are useful," (Box, 1979) should be understood. Explainable AI generates approximate simple models and calls them 'explanations', suggesting reliable knowledge of how a complex model functions.

When characterising the use of such models in science, Hesse (1965) divides the properties of the model into positive analogies, where the properties of the model are known to correspond to properties of the phenomena we are interested in; negative analogies, where the properties of the model do not match the phenomena we are interested in; and neutral analogies, where it is unknown if the properties of the model correspond to the phenomena.

This characterisation captures many of the challenges when offering approximations of models as explanations. It is not enough to simply offer a human interpretable model as an explanation. For an individual to be able to trust such a model as an approximation, they must know over which domain a model is reliable and accurate, where it breaks down, and where its behaviour is uncertain. If the recipient of a local approximation does not understand its limitations, at best it is not comprehensible, and at worst misleading.

This is not to say that local approximations are without merit, but rather that they can only reliably have explanatory power if

[5]Famously, GPS satellites are insufficiently accurate unless they account for effects of general relativity.

their limitations are clearly documented and understood by recipients. For domain experts with in-depth knowledge of when and where approximations break down, or for technicians that have a clearly defined and tested remit for where the approximation can be used, they can be extremely useful. However, at the moment xAI generally avoids the challenges of testing and validating approximation models, or fully characterising their domain. If these elements are well understood by the individual, models can offer more information than an explanation of a single decision or event. Over the domain for which the model accurately maps onto the phenomena we are interested in, it can be used to answer 'what if' questions, for example "What would the outcome be if the data looked like this instead?" and to search for contrastive explanations, for example "How could I alter the data to get outcome X?"

However, local models may also provide false assurances. As suggested above, local approximations are often misleading or inaccurate outside of their domain, and provide little insight into how a function response and outcomes vary with changes to the inputs. By definition local explanations hold only for a specific decision; what is explained is not how the model functions as a whole, but rather one segment of the model relevant to the prediction at hand. Thus, while helpful to explain the weights and relationships between variables in a small segment of a model (relevant to a particular case or decision), the explanations do not provide evidence of the trustworthiness or acceptability of the model overall.

3.2 Linear approximations

To provide further support for the analogue between xAI and scientific modelling, we turn now to variants of linear approximation. Many of the issues and design decisions made by researchers in the field point directly towards the issues with modelling previously, particularly the three-way trade-off between the simplicity of the approximated model, the size of the domain it describes, and the accuracy of this description.

3.2.1 Linear Models in Continuous Spaces. Linear models are designed to give a single measure of the importance of each variable to the classifier they are approximating. In some cases (e.g. Lundberg and Lee (2017); Ribeiro et al. (2016), these weights can be directly interpreted as the sensitivity or a number that tells you how much the classifier response will vary as a particular feature changes. In other words, if a particular variable has a weight of associated with it, altering the variable by small amount will cause the classifier to vary by approximately . Regardless of whether linear models are intended to be a local approximation of a classifier or a simple importance measure, they suffer from two distinct issues. The first is a problem of curvature, namely that the sensitivity of a classifier to a change in a particular variable may vary with the amount the variable changes. The second issue is one of dependencies between variables and how to capture the relationships between them.

Both of these issues can be illustrated by an example taken from Miller (2017). Suppose, I believe that a particular creature is a bee because I have been told it has 6 legs and four wings. I am unlikely to change my mind if I am told it actually has three wings. The most reasonable explanation for a three winged insect is that it was originally a four winged insect that lost a wing, but if I'm told that it has two wings I may well believe that it is a fly. This is an example of the sensitivity varying with the size of the change made. On the other hand, I will only be comfortable believing it is a spider if it has both eight legs and no wings. Should I then say that my belief depends wholly on the number of legs it has; wholly on the number of wings; or weakly on both? As linear models do not capture the interdependencies between variables, they cannot accurately characterise these relationships. In all such cases, when a local model is fitted to the classifier response, the choice of domain or variable values governs how well the approximation performs.

3.2.2 Gradient Sensitivity verses Binarization. There are two main schools of thought regarding the problems of varying sensitivity and scale. The first notes that the sensitivity of a classifier is only well defined when the size of in the previous section tends towards zero. In this case, the sensitivity is equivalent to the gradient. Although well defined, this essentially means that the model is fitted to a domain of size 0 and may exhibit large amounts of instability while offering limited predictive power. Two prominent approaches are described by Simonyan et al. (2013) and Baehrens et al. (2010).

The other option, and most influentially proposed by Ribeiro et al. (2016) in their Locally Interpretable Model-Agnostic Explanations (LIME) approach, is to binarize the problem. Rather than trying to fit a linear classifier to a large range of values, the authors consider a binary problem, where for each feature they attempt to switch it on and off, allowing them to answer the question "What is the contribution of feature f to the classifier response, given the data it currently sees?" This leaves open the question of "the contribution compared to what?" For unstructured data, such as a count of how many times particular words occur in a document, it makes sense to compare against a baseline created by setting the count to 0. For structured data this is more problematic. For example, how can we evaluate the importance of someone's salary to a loan decision, if the classifier can only evaluate people with valid salaries? The answer is to compare it against a different valid salary, but it is unclear how this valid salary should be chosen.

This issue is even more apparent when creating local approximations of computer vision algorithms, as individual pixels cannot be removed from an image, but only set to different values. Several options have been proposed. LIME appeared to set regions of the image to an unspecified homogeneous value (Ribeiro et al., 2016). Deep Taylor Decomposition (Montavon et al., 2017) suggests blurring the image, an operation which preserves colour information but removes texture. DeepLift uses a user-specified value (Shrikumar et al., 2017, 2016), while layerwise relevance propagation sets internal network values to 0 (Montavon et al., 2017). Each of these choices is an implicit restriction of the domain over which the model is fitted and carries different implications for the kinds of model found, and can substantially alter the importance given to features. For example, contrasting current image values against a particular colour, such as grey, makes it appear that grey pixels have no effect, while contrasting an image against its blurred version makes it appear as if only high-frequency texture cues are important and that the classifier does not use colour information.

3.2.3 Linear models in high-dimensional spaces. Having made a choice of a binarization as discussed in the previous section, a question then remains as to which of these values to approximate

with a linear model. Even after restricting a function defined over a continuous high-dimensional range to a set of binary variables, this gives a (hyper-)cube of possible values while a linear function can only uniquely specify values, with all other values being a linear approximation of that (see Figure 1 for an illustration). This raises the question of which of these values are important to approximate. For example, should we only pay attention to contrasting solutions closest to the original solution as suggested by DeepLift (illustrated by Figure 1 centre; (Shrikumar et al., 2017, 2016)); uniformly weight over all possible values, suggested by SHAPE (Lundberg and Lee, 2017) (Figure 1 right); or pay as much attention to values close to the data point as to those close to the alternate solution (e.g. a solid grey image); or a weighted mixture of the two approaches (e.g. Figure 1 left LIME; (Ribeiro et al., 2016))?

3.3 Exploring alternatives to scientific modelling

Local approximations thus face difficulties with generalizability, arbitrariness in choice of domain, and the potential to mislead recipients unless the domain and epistemic limitations of the approximation are known. Given these difficulties, other methods of producing explanations may be preferable from the perspective of the user or individual affected by a black-box system. The utility of local approximations is dependent upon the knowledge of the recipient regarding the approximations limitations, including conditions under which it will break down and provide a misleading explanation of the decision-making model. Local approximations can be useful as a type of 'explanation kit' or causal chain that allows expert users to explore slices of a model for prototyping or debugging (Miller, 2017, p.17)(Poulin et al., 2006). However, their ultimate utility and reliability for non-experts, including individuals subject to decisions made by the system, is highly questionable.

This finding raises the question: might other methods for generating explanations perform better, or at least offer different benefits, than local approximations? Can other methods provide more reliable or personally relevant information for non-experts to enhance accountability and trustworthiness in algorithmic systems?

4 CONTRASTIVE EXPLANATIONS

Thus far we have discussed the philosophical and practical purposes of scientific models, which can be understood as partial causal scientific explanations that assist in comprehending a piece of the functionality of a phenomenon (Ruben, 2004). Given the difficulties faced by local approximations, it is worth examining prior work in the 'explanation sciences' to identify potential alternative approaches to generate reliable and practically useful post-hoc interpretations for parties affected by an algorithmic decisions. If our goal is to produce explanations that are comprehensible and useful to expert as well as non-expert stakeholders, it is sensible to examine theoretical as well as empirical work describing how humans give and receive explanations (Miller, 2017, p.3-4).

In recent decades, work in the philosophy of science and epistemology has paid increasing attention to theories of contrastive explanations and counterfactual causality (e.g. Kment (2006); Lewis (1973); Ruben (2004); Woodward and Zalta (2003)). In short, contrastive theories argue that causal explanations inevitably involve appeal to a counterfactual case, be it a cause or event, which did not occur. A canonical example is provided by Lipton Lipton (1990): "To explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the history of not-Q" . Some authors go so far as to claim all questions about causality are inherently contrastive (Lewis, 1973; Ruben, 2004).

Contrastive theories of explanation are of course not without criticism. Ruben (2004), for example, has suggested that, even if causal explanations are inevitably contrastive in nature (which he doubts), this characteristic can be dealt with by traditional theories of explanation, rendering the 'contrastive turn' interesting but ultimately unnecessary. While the utility of contrastive theories remains debated, the fact that contrastive explanations address a particular event or case and are thus simpler to generate than complete or global explanations of model functionality suggest they worth further consideration in xAI (Lipton, 1990).

A recent review by Miller (2017) suggests substantial empirical support exists for the practical utility of 'everyday' contrastive explanations. Miller reviewed articles and empirical studies from "philosophy, psychology, and cognitive science of how people define, select, evaluate, and present explanations" (Miller, 2017, p.1). His analysis highlighted three primary characteristics of explanations as they are used, selected, evaluated, and shared by individuals:

4.1 Human explanations are contrastive

'Everyday explanations' are "sought in response to particular counterfactual cases...That is, people do not ask why event P happened, but rather why event P happened instead of some event Q" (Miller, 2017, p.5).[6] The preference for contrastive explanations is not due merely to the cognitive complexity of non-contrastive explanations, for example the number of links in a causal chain. Rather, the reviewed empirical evidence indicates that humans psychologically prefer contrastive explanations (Miller, 2017, p.18)(Rehder, 2003, 2006).

The perceived abnormality of an event influences requests for contrastive explanations which address why a normal or expected event did not occur (Hilton and Slugoski, 1986; McClure et al., 2003; Samland and Waldmann, 2014). 'Normal' behaviour has empirically been shown to be judged as "more explainable than abnormal behaviour," with perceived abnormality playing an important role in explanation selection (Miller, 2017, p.41). Gregor and Benbasat (1999) support the importance of abnormality, suggesting that users request explanations when an anomaly or abnormal event is detected. Lim and Dey (2009) similarly note a positive relationship between the perceived "inappropriateness" of application behaviour and user requests for contrastive explanations. Violation of ethical and social norms can likewise set an event apart as abnormal (Hilton, 1996). Explanations addressing why an alternative, expected event

[6] It is worth noting that counterfactual cases in contrastive, everyday explanations of algorithmic decisions are not equivalent to counterfactuals used to assess causality (Miller, 2017, p.13)(Hilton and Slugoski, 1986; Woodward, 1997), in the sense that the range of possible alternatives is necessarily bounded by the limitations of the model in question and the features available to it. This invariance in the model allows for reliable contrastive or counterfactual explanations to be computed (Woodward, 1997).

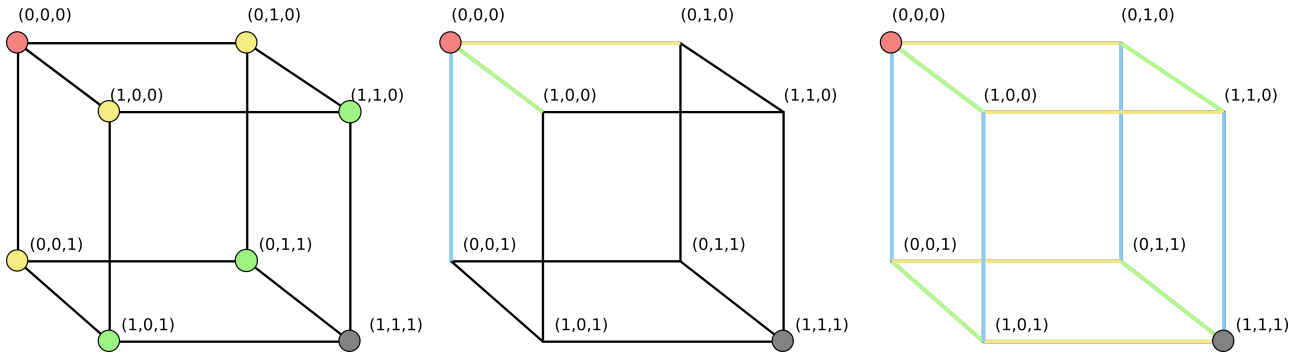

Figure 1: An illustration of the different weighting scheme used in fitting linear models. LIME (leftmost) weights all examples differently based on how far they are from the original data point (illustrated by different coloured vertices), while DeepLift (centre) only fits the linear weights to the individual edges closest to the original data point (coloured pink). SHAPE (right) fits weights by averaging over all edges formed by flipping a single variable from off to on (each group averaged together is indicated by a single colour). Each of these different approaches equates to a different assumption as to which samples are most important, and none of them can be said *a priori* to be better than any of the others.

did not occur have historically been addressed in the experts systems literature, seen for instance in the discussion of "Why Not" explanations by Lim and Dey (2009).

4.2 Human explanations are selective

Full or scientific explanations are rarely if ever realised in practice. Absent general laws or the complete causal chain leading to an event, multiple explanations are typically possible that attribute different causes. A given cause of set of causes may be incomplete insofar as they are not the sole cause of the event, but nonetheless convey useful information to the recipient in a given context or for a given purpose (Ylikoski, 2013). As Miller argues, "Explanations are selected – people rarely, if ever, expect an explanation that consists of an actual and complete cause of an event. Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be the explanation" (Miller, 2017, p.5). Further, to be informative, explanations should not be entirely reducible to presuppositions, or beliefs that the recipient of the explanation already holds (Hesslow, 1988). They should further be relevant to the question asked by the recipient (epistemic relevance; (Miller, 2017), or what Slugoski et al. (1993) describe as the recipient's context.

When an explanation giver ('explainer') selects an explanation for an event, possible or actual causes can be 'backgrounded' or 'discounted', meaning they are disregarded on the basis of contextual information that renders them irrelevant to the purposes of the explainer or recipient of an explanation (the 'explainee'). This type of selection is essential to reduce long causal chains to a cognitively manageable size (Hilton, 1996). In xAI, selection often takes the form of key features or evidence being emphasised in explanation interfaces based upon their relative weight or influence on a given prediction or output (Biran and McKeown, 2014; Poulin et al., 2006). As the observations above suggest, the relevance of features (and explanations addressing them) would be based not only on 'statistical weight', but also the explainee's subjective interests and expectations.

4.3 Human explanations are social

Explanations are social, insofar as they involve an interaction between one or more explainers and explainees. Interactive transfer of knowledge is required in which information is tailored according to the recipient's beliefs and comprehensional capacities (Miller, 2017, p.5). Explanations can be conceived as involving one or more explainers and explainees engaging in information transfer through dialogue, visual representation, or other means (Hilton, 1990), often to correct information or knowledge assymetry (Lim and Dey, 2009). In the case of machine learning models, it is perhaps most useful to always treat explanation generation as an interactive process, initially involving a mix of human and automated actors, at a minimum an inquirer (e.g. a developer, user) and the model or system (Kayande et al., 2009; Martens and Provost, 2013). Further, explanations are iterative, insofar as they must be selected and evaluated on the basis of shared presuppositions and beliefs. Relevance is key, and iteration may be required to communicate effectively or clarify points of confusion on the path towards a mutually understood explanation.

Together, these characteristics of everyday explanations reveal that they "are not just the presentation of causes (causal attribution). While an event may have many causes, often the explainee cares only about a small subset (relevant to the contrast case), the explainer selects a subset of this subset (based on several different criteria), and explainer and explainee may interact and argue about this explanation" (Miller, 2017, p.6).

4.4 Contrastive explanations in xAI

Contrastive methods of generating explanations are responsive to these three characteristics of explanations emphasised in the 'explanation sciences'. Two approaches for directly computing contrastive explanations are described by Martens and Provost (2013) and Wachter et al. (2018).[7] Such post-hoc methods avoid many

[7] These methods resemble the aforementioned "Why Not" explanations described by Lim and Dey (2009), and are related to work on adversarial perturbations (e.g. (Dube, 2018; Goodfellow et al., 2014)

of the difficulties faced by model-based explanations. Rather than explicitly generating a model that approximates functional values over a restrictive domain, and relying on the user to interpret this, contrastive explanations directly offer an alternative data point: "If your data had looked like this, you would have been given this classification score instead." These alternative data points can be computed exactly. As such, many of the challenges facing 'modelling' approaches to generating explanations, such as the quality of the approximation or the limits of a chosen domain, do not arise to a comparable degree.

However, as contrastive methods only return a single data point, a more pressing concern is the relevance of the output. If this data point does not directly correspond to a factoid of interest to the user, it cannot be used to deduce relevant conclusions, for example regarding the justifiability of a decision. Similar issues arise in the fitting of models: if the domain the model is fitted to does not capture examples relevant to the intended audience, then it is unlikely to be useful.

The first approach described by Martens and Provost (2013) is explicitly designed for use on discrete data, and focuses on the particular problem of which words need to be removed from a website in order for it to be no longer be classified as an adult (i.e. pornographic) website. In contrast, Wachter et al. (2018) propose a method, 'counterfactual explanations', designed to work on primarily continuous data. They illustrate their approach on the problem of law school admissions and risk factors likely to increase a patient's chance of developing diabetes.

Finding such counterfactuals explanations can be described as a search or optimisation problem. Such approaches seek a similar counterfactual that is both close to the original datapoint and likely to occur in the real world (Kment, 2006). In the case of Wachter et al. (2018) this was formulated as a Lagrangian style constrained optimisation:

$$\arg\min_{\mathbf{c}} \max_{\lambda} \lambda(f(\mathbf{c}) - T)^2 + d(\mathbf{c}, \mathbf{x}) \tag{1}$$

Where $\mathbf{x}$ is the original data point, and $\mathbf{c}$ the counterfactual. $f(\mathbf{c})$ is the classification response from the black-box function f, which is constrained to take target value T. $d(\cdot, \cdot)$ is a distance function that ensures that the counterfactual is a relevant, and a human comprehensible change to the original datapoint.

5 TOWARDS COMMUNICATIVE, CONTRASTIVE EXPLANATIONS

Such methods for computing contrastive explanations seek to provide contextually-relevant information to parties affected by a decision by describing how relevant closely related, alternative events could have occurred. However, choosing a relevant set of cases or events against which contrastive explanations are provided is not a straightforward challenge. The way in which information is transferred has a substantial impact on the quality and psychological acceptability of explanations (Hilton, 1990). The recipient's beliefs about an event to be explained are similarly constrained by the explainer's choice of explanation. As a result, the explainer's epistemological and normative values can have a significant effect on the recipient's understanding of an event. Lombrozo (2009) for example demonstrated that the type of explanation provided (e.g. mechanistic, functional) can influence the recipient's view of the importance of features in categorising a phenomenon (Poulin et al., 2006). Explainers and explainees can have different motivations, for example to generate trust (explainer) or understand non-intuitive causes of a decision (explainee), meaning conflicts can arise in explanation selection and evaluation.

The normativity of this relationship means that a risk exists of malicious explainers subtly discouraging explainees from critically questioning or contesting a decision through choice of explanation(s). This risk is particularly acute when explanations are given to foster trust or understanding. A recipient's beliefs can potentially be 'gamed' or manipulated to align with the explainer's preferred explanation of a phenomenon, meaning recipients can be 'nudged' to take a preferred action (or not). Seemingly rational bases for otherwise unjustifiable decisions can for example be offered to nudge the recipient not to question or contest the decision. Lipton (2016) has urged for caution in adopting post-hoc interpretation methods due to this potential to mislead recipients, for example by falsely attributing a decision to an irrelevant feature, or a more acceptable feature (e.g. post code, 'leadership') that serves as a proxy for a legally protected feature (cf. Barocas and Selbst (2016); Kim (2016)).

These risks may be mitigated by developing methods to provide contrastive, selective, and social explanations that not only enable information exchange and dialogue between giver and recipient, but critical argumentation and discussion of the justifiability of an event as well. This suggestion stems from prior work that suggests conversational explanations are essentially a form of argumentation. Conversational explanations are used to both offer causes of an event and back claims as to the truth or relevance of these causes (Antaki and Leudar, 1992). Walton (2004, 2007) takes a similar approach in proposing a dialectical theory of everyday explanations, which suggests that giving explanations involves not only transfer of information or causal claims, but also argumentative support for these claims.

These findings, that explanations are given via conversation and resemble argumentation, reveal a mechanistic link between explanation and justification as a type of discourse (Fox et al., 2007). Justification is also often a discursive act, relying upon explanations to transfer knowledge and support claims made by each party. Interest in justification stems from an individual's ability to comprehend decisions made about them, and contest them when they are obviously incorrect or found unacceptable (or unjustifiable). As such, argumentation models of explanation resemble theories of justification and democracy based upon discourse, such as Jurgen Habermas' discourse ethics (located in his broader Theory of Communicative Action; (Habermas, 1984)).

Justification has recently received increasing attention by scholars discussing algorithmic accountability (e.g. Binns (2017); Biran and McKeown (2014); Hildebrandt and Koops (2010); Selbst and Barocas (2018); Weller (2017)). However, justification currently occupies an uneasy position in xAI, in that it is rarely formally defined or related to explanations or ideals of transparency and accountability.[8] This unease is perhaps understandable; as Binns (2017) notes, legitimate disagreements between epistemic and ethical standards

[8] Biran and McKeown (2014) prove an exception to this rule. They define explanation as an answer to the question "how did the system arrive at the prediction?" whereas justification answers the question "why should we believe the prediction is correct?"

for algorithmic decision-making can exist which require resolution, for example through democratic processes. Despite this, important conversations around the ethical acceptability or justifiability of algorithmic systems must occur in society if responsible deployment is sought.

Finally, very little work in xAI addresses the link between interpretability and contestability of models or decisions (Lipton, 2016). Going forward, explanations of specific algorithmic decisions should allow the justification of a black-box model or decision to be debated and contested.. If we use justificatory discourse as a framework for explanation requirements, we need to determine what sort of records algorithmic systems must retain in order to allow for contesting and post-hoc auditing of abnormal events (Mittelstadt, 2016; Sandvig et al., 2014), for instance through identification of classification errors or inaccurate input data (Poulin et al., 2006). Where approximations are used to provide explanations, information should also be provided to affected parties detailing the limitations and relative resilience of the approximation, including the domain addressed and why it has been chosen. Further, meaningful, critical dialogue can be achieved between user, developer, and model by ensuring explanations are contrastive, selective, and social. We thus need to ensure xAI aims to develop methods for producing explanations of model functionality and specific decisions that embody these characteristics. Reliable methods for producing contrastive explanations and explanation by approximation are both required.

At the moment, the xAI community largely fails in this task. Many approaches produce approximate and local models that are more akin to models in science. It may be possible to ask questions of these models, and to develop contrastive explanations from them, and so they are not without value. But they do not help us directly provide contrastive explanations to parties affected by algorithmic decisions. Going forward, the field must urgently close this gap.

REFERENCES

Philip Adler, Casey Falk, Sorelle A. Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2016. Auditing Black-box Models by Obscuring Features. *arXiv:1602.07043 [cs, stat]* (22 2 2016). http://arxiv.org/abs/1602.07043 arXiv: 1602.07043.

Charles Antaki and Ivan Leudar. 1992. Explaining in conversation: Towards an argument model. *European Journal of Social Psychology* 22, 2 (1992), 181–194.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11, Jun (2010), 1803 – 1831.

Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104, 3 (2016). https://doi.org/10.15779/Z38BG31

Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpretability via Model Extraction. *arXiv:1706.09773 [cs, stat]* (29 6 2017). http://arxiv.org/abs/1706.09773 arXiv: 1706.09773.

Bettina Berendt and Sören Preibusch. 2017. Toward Accountable Discrimination-Aware Data Mining: The Importance of Keeping the Human in the Loop and Under the Looking Glass. *Big Data* 5, 2 (1 6 2017), 135–152. https://doi.org/10.1089/big.2016.0055

Reuben Binns. 2017. Algorithmic Accountability and Public Reason. *Philosophy & Technology* (24 5 2017), 1–14. https://doi.org/10.1007/s13347-017-0263-5

Or Biran and Kathleen McKeown. 2014. Justification narratives for individual classifications. In *Proceedings of the AutoML workshop at ICML*, Vol. 2014.

B. Bodo, N. Helberger, K. Irion, F. Zuiderveen Borgesius, J. Moller, B. van de Velde, N. Bol, B. van Es, and C. de Vreese. 2017. Tackling the algorithmic control crisis-the technical, legal, and ethical challenges of research into algorithmic agents. *Yale JL & Tech.* 19 (2017), 133.

George EP Box. 1979. Robustness in the strategy of scientific model building. In *Robustness in statistics*. Elsevier, 201–236.

Jenna Burrell. 2016. How the Machine 'Thinks:' Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* (2016). https://doi.org/10.1177/2053951715622512

R. Caruana, H. Kangarloo, J. D. Dionisio, U. Sinha, and D. Johnson. 1999. Case-based explanation of non-case-based learning methods. *Proceedings of the AMIA Symposium* (1999), 212–215. PMID: 10566351 PMCID: PMC2232607.

Junxiang Chen, Yale Chang, Brian Hobbs, Peter Castaldi, Michael Cho, Edwin Silverman, and Jennifer Dy. 2016. Interpretable Clustering via Discriminative Rectangle Mixture Model. *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, 823 – 828. http://ieeexplore.ieee.org/abstract/document/7837910/ [Online; accessed 2017-10-16].

Danielle Keats Citron and Frank Pasquale. 2014. The scored society: due process for automated predictions. *Wash. L. Rev.* 89 (2014), 1.

Mark Craven and Jude W. Shavlik. 1996. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, 24 – 30. http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf [Online; accessed 2017-10-16].

Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. IEEE, 598–617. https://doi.org/10.1109/SP.2016.42 [Online; accessed 2016-09-12].

Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134* (2017).

Simant Dube. 2018. High Dimensional Spaces, Deep Learning and Adversarial Examples. *arXiv preprint arXiv:1801.00634* (2018).

Ruth C. Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. *arXiv preprint arXiv:1704.03296* (2017).

John Fox, David Glasspool, Dan Grecu, Sanjay Modgil, Matthew South, and Vivek Patkar. 2007. Argumentation-based inference and decision making–A medical perspective. *IEEE intelligent systems* 22, 6 (2007).

They further suggest that a satisfactory answer to the first question will also provide an answer to the second for explainees with sufficient expertise.

Roman Frigg. 2006. Scientific Representation and the Semantic View of Theories. *Theoria* 55 (2006), 37–53.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly* (1999), 497–530.

J. Habermas. 1984. *The Theory of Communicative Action: Volume 1: Reason and the Rationalization of Society*. Beacon, Boston.

William Herfel, Wladiyslaw Krajewski, Ilkka Niiniluoto, and Wojcicki (Eds.). 1995. *Theories and Models in Scientific Process*. Rodopi, Amsterdam.

Mary B. Hesse. 1965. Models and analogies in science. (1965).

Germund Hesslow. 1988. The problem of causal selection. *Contemporary science and natural explanation: Commonsense conceptions of causality* (1988), 11–32.

Mireille Hildebrandt and Bert-Jaap Koops. 2010. The Challenges of Ambient Law and Legal Protection in the Profiling Era. *The Modern Law Review* 73, 3 (May 2010), 428–460. https://doi.org/10.1111/j.1468-2230.2010.00806.x

Denis J. Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65.

Denis J. Hilton. 1996. Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning* 2, 4 (1996), 273–308.

Denis J. Hilton and Ben R. Slugoski. 1986. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological review* 93, 1 (1986), 75.

Ujwal Kayande, Arnaud De Bruyn, Gary L. Lilien, Arvind Rangaswamy, and Gerrit H. Van Bruggen. 2009. How incorporating feedback mechanisms in a DSS affects DSS evaluations. *Information Systems Research* 20, 4 (2009), 527–546.

Been Kim, Cynthia Rudin, and Julie A. Shah. 2014. The bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in Neural Information Processing Systems*, 1952–1960.

Pauline T. Kim. 2016. Data-driven discrimination at work. *Wm. & Mary L. Rev.* 58 (2016), 857.

Boris Kment. 2006. Counterfactuals and explanation. *Mind* 115, 458 (2006), 261–310.

Joshua A. Kroll, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2016. Accountable algorithms. *U. Pa. L. Rev.* 165 (2016), 633.

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. ACM Press, 126–137. https://doi.org/10.1145/2678025.2701399 [Online; accessed 2018-05-06].

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & Explorable Approximations of Black Box Models. *arXiv:1707.01154 [cs]* (4 7 2017). http://arxiv.org/abs/1707.01154 arXiv: 1707.01154.

Bruno Lepri, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. 2017. Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy & Technology* (15 8 2017). https://doi.org/10.1007/s13347-017-0279-x [Online; accessed 2017-08-25].

David Lewis. 1973. *Counterfactuals*. Blackwell, Oxford.

Brian Y. Lim and Anind K. Dey. 2009. Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing - Ubicomp '09*. ACM Press, Orlando, Florida, USA, 195. https://doi.org/10.1145/1620545.1620576

Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplements* 27 (1990), 247–266.

Zachary C. Lipton. 2016. The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]* (10 6 2016). http://arxiv.org/abs/1606.03490 arXiv: 1606.03490.

Paulo JG Lisboa. 2013. Interpretability in Machine Learning – Principles and Practice. In *Fuzzy Logic and Applications*. Springer, 15 – 21. http://link.springer.com/chapter/10.1007/978-3-319-03200-9_2 [Online; accessed 2015-12-19].

Tania Lombrozo. 2009. Explanation and categorization: How "why?" informs "what?". *Cognition* 110, 2 (2009), 248–253.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. 4765–4774.

Alessandro Mantelero. 2016. Personal data for decisional purposes in the age of analytics: From an individual to a collective dimension of data protection. *Computer law & security review* 32, 2 (2016), 238–255.

David Martens, Bart Baesens, Tony Van Gestel, and Jan Vanthienen. 2007. Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research* 183, 3 (2007), 1466–1476.

David Martens and Foster Provost. 2013. Explaining data-driven document classifications. (2013). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2282998 [Online; accessed 2017-09-22].

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. ACM Press, 165–172. https://doi.org/10.1145/2507157.2507163 [Online; accessed 2017-09-24].

John L. McClure, Robbie M. Sutton, and Denis J. Hilton. 2003. The Role of Goal-Based Explanations. *Social judgments: Implicit and explicit processes* 5 (2003).

Tim Miller. 2017. Explanation in artificial intelligence: Insights from the social sciences. *arXiv preprint arXiv:1706.07269* (2017).

Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *arXiv:1712.00547 [cs]* (1 12 2017). http://arxiv.org/abs/1712.00547 arXiv: 1712.00547.

Brent Mittelstadt. 2016. Automation, Algorithms, and Politics| Auditing for Transparency in Content Personalization Systems. *International Journal of Communication* 10 (2016), 12.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2017. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* (2017).

Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and engineering ethics* 2, 1 (1996), 25–42.

S. C. Olhede and P. J. Wolfe. 2018. The growing ubiquity of algorithms in society: implications, impacts and innovations. *Phil. Trans. R. Soc. A* 376, 2128 (Sept. 2018), 20170364. https:

//doi.org/10.1098/rsta.2017.0364

Frank Pasquale. 2015. *The black box society: The secret algorithms that control money and information.* Harvard University Press.

Brett Poulin, Roman Eisner, Duane Szafron, Paul Lu, Russ Greiner, D S Wishart, Alona Fyshe, Brandon Pearcy, Cam MacDonell, and John Anvik. 2006. Visual Explanation of Evidence in Additive Classifiers. (2006), 8.

Bob Rehder. 2003. A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29, 6 (2003), 1141.

Bob Rehder. 2006. When similarity and causality compete in category-based property generalization. *Memory & Cognition* 34, 1 (2006), 3–16.

Marco Tulio 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*. ACM Press, 1135–1144. https://doi.org/10.1145/2939672.2939778 [Online; accessed 2017-09-24].

David-Hillel Ruben. 2004. *Explaining explanation.* Routledge.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv preprint arXiv:1708.08296* (2017). https://arxiv.org/abs/1708.08296 [Online; accessed 2017-09-22].

Jana Samland and Michael R. Waldmann. 2014. Do Social Norms Influence Causal Inferences? *Proceedings of the Annual Meeting of the Cognitive Science Society* 36.

Ivan Sanchez, Tim Rocktaschel, Sebastian Riedel, and Sameer Singh. 2015. Towards extracting faithful and descriptive representations of latent variable models. *AAAI Spring Syposium on Knowledge Representation and Reasoning (KRR): Integrating Symbolic and Neural Approaches* (2015). http://www.aaai.org/ocs/index.php/SSS/SSS15/paper/viewFile/10304/10033 [Online; accessed 2017-10-16].

Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* (2014), 1–23.

Andrew D. Selbst and Solon Barocas. 2018. The intuitive appeal of explainable machines. (2018).

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *See https://arxiv.org/abs/1610.02391 v3* (2016). https://pdfs.semanticscholar.org/5582/bebed97947a41e3ddd9bd1f284b73f1648c2.pdf [Online; accessed 2017-10-16].

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. CoRR.

Avanti Shrikumar, Peyton Greenside, and Anna Shcherbina. 2016. Not just a black box: Learning important features through propagating activation differences. CoRR.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013). https://arxiv.org/abs/1312.6034 [Online; accessed 2017-09-24].

Ben R. Slugoski, Mansur Lalljee, Roger Lamb, and Gerald P. Ginsburg. 1993. Attribution in conversational context: Effect of mutual knowledge on explanation–giving. *European Journal of Social Psychology* 23, 3 (1993), 219–238.

Paolo Tamagnini, Josua Krause, Aritra Dasgupta, and Enrico Bertini. 2017. Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations. ACM Press, 1–6. https://doi.org/10.1145/3077257.3077260 [Online; accessed 2017-09-22].

Michael Veale and Lilian Edwards. 2018. Clarity, surprises, and further questions in the Article 29 Working Party draft guidance on automated decision-making and profiling. *Computer Law & Security Review* 34, 2 (2018), 398–404.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology* forthcoming (2018).

Douglas Walton. 2004. A new dialectical theory of explanation. *Philosophical Explorations* 7, 1 (2004), 71–89.

Douglas Walton. 2007. Dialogical Models of Explanation. *ExaCt* 2007 (2007), 1–9.

Fulton Wang and Cynthia Rudin. 2015. Falling rule lists. *Artificial Intelligence and Statistics*, 1013–1022.

Adrian Weller. 2017. Challenges for Transparency. *arXiv:1708.01870 [cs]* (29 7 2017). http://arxiv.org/abs/1708.01870 arXiv: 1708.01870.

Jim Woodward. 1997. Explanation, Invariance, and Intervention. *Philosophy of Science* 64 (1997), S26–S41. https://www.jstor.org/stable/188387

James Woodward and E. Zalta. 2003. *Scientific explanation.*

Petri Ylikoski. 2013. Causal and constitutive explanation compared. *Erkenntnis* 78, 2 (2013), 277–297.

Tal Z. Zarsky. 2013. Transparent predictions. *U. Ill. L. Rev.* (2013), 1503.

Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180, 3 (2017), 689 – 722.