Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1780–1790
Marseille, 11–16 May 2020

Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models

Christopher Grimsley[1]*, Elijah Mayfield[2]*, and Julia R.S. Bursten[1]
[1]Department of Philosophy, University of Kentucky
[2]Language Technologies Institute, Carnegie Mellon University
christopher.grimsley@uky.edu, elijah@cmu.edu, jrbursten@uky.edu

Abstract

As the demand for explainable deep learning grows in the evaluation of language technologies, the value of a principled grounding for those explanations grows as well. Here we study the state-of-the-art in explanation for neural models for NLP tasks from the viewpoint of philosophy of science. We focus on recent evaluation work that finds brittleness in explanations obtained through attention mechanisms. We harness philosophical accounts of explanation to suggest broader conclusions from these studies. From this analysis, we assert the impossibility of causal explanations from attention layers over text data. We then introduce NLP researchers to contemporary philosophy of science theories that allow robust yet *non-causal* reasoning in explanation, giving computer scientists a vocabulary for future research.

Keywords: philosophy of science, explainability, causal reasoning, attention mechanisms.
* denotes equal contributions.

1. Introduction

In the natural language processing community, we've reached a consensus that *explainability* in trained models is a positive attribute. When performing model selection, the less complex and more explainable model should be preferred (holding all else - for instance, classification accuracy or training time - equal). Part of this is purely intuitive and based on logistic ease for software developers; the preference for explainable models is also spurred on by regulation, led by the European Union's "right to explanation" in the 2016 enactment of the GDPR (Goodman and Flaxman, 2017). Yet so far, most researchers in NLP that describe their models as explainable have treated explanation the way that U.S Supreme Court Justice Potter Stewart famously treated obscenity: *"I know it when I see it"* (Stewart, *concurring*, 1964). It is challenging to evaluate the success of an explainable neural model without defining a criterion for evaluating what counts as an *explanation*, and moreover what counts as a *good* explanation.

This may be because the phenomenon of giving, receiving, and incorporating explanations is like the breathed air of the process of scientific reasoning: explanation is how knowledge is transmitted, how mechanisms and causal processes are learned, how experimental design problems are solved, and how publications and presentations are assembled. Because it appears everywhere, its value and form is assumed. However, like air, explanation can vary dramatically in different situations, and assuming that its structure and function in one setting can be automatically exported in another can be as dangerous as mistaking oxygen for nitrogen.

Research in philosophy of science has produced decades of results on the nature of scientific explanations across fields, and related those findings to case studies across scientific disciplines. In the NLP community, some of this research has been anticipated by, and recapitulated in, parallel work on explainable neural networks. But many of the distinctions that philosophers have discovered have gone unnoticed in our field. Even when machine learning publications aim to characterize or define explanation by distinguishing it from other logical and psychological phenomena, their authors tend to begin by assuming that there is some singular thing that is an explanation, and that other researchers will *"know it when they see it"* in the same way that the paper's authors do. Moreover, they assume that explainability can be measured against other measurable quantities in machine learning models, in the context of tradeoffs and holistic assessment of model quality. There is a long list of explanainable AI overview articles proposing desirable characteristics of what an explanation should look like (Lipton, 2016). But a central contention of philosophy of science is that no such obvious and univocal phenomenon exists.

This paper seeks to introduce contemporary philosophy of science debates to the explainable NLP community, and we make several contributions to build a theoretical grounding for use in explanation of deep neural models. First, in section 2. we walk through the progression of explanations in machine learning, stepping from rule-based systems to linear models and most recently to deep models. We end that section with a description of one widespread and highly popular approach for explaining neural networks, particularly in tasks using text or speech inputs: the use of *attention mechanisms* as a functional basis for generating explanations. We then give a detailed summary of two recent findings on the limitations of attention mechanisms for explainable NLP.

1. Jain and Wallace (2019), which finds that attention layers in neural networks can be subject to adversarial reweightings, undermining their use for explanation.

2. Serrano and Smith (2019), which finds that a very large number of attention weights can be zeroed out entirely, again undermining the use of these layers for identifying the importance of intermediate representations within a deep neural classifier.

We follow this with an overview of philosophical theories of scientific explanation in section 3., including brief sum-

maries of multiple competing and complementary perspectives. This overview is itself a new contribution, both in the taxonomy of theories it develops and in its presentation for computer scientists in the explainable NLP community. Next, in section 4. we provide a more robust introduction of one theory of explanation in particular, the *interventionist* account of causal explanation. Due to its emphasis on causal reasoning via counterfactual analysis and its historical development at the interface of computer science and philosophy of science, this account is a good tool for analyzing a class of recent findings on explanation of neural networks through causal reasoning. With an established background from both philosophy of science and deep learning, in section 5. we apply the interventionist account to the study of attention mechanisms for explaining neural network behavior. We focus on three primary research questions:

Q1. Is it appropriate to analyze these studies using the interventionist account?

Q2. The account requires that interventions be surgical in order to make causal claims. Do the studies succeed at surgical intervention?

Q3. If attention weights cannot be manipulated surgically, what are the consequences for explanation through attention?

Yet the interventionist account deals only in *causal* explanation in the sciences. An important corollary of our analysis is that when a network will not and cannot produce causal explanations, it can still render alternate, non-causal types of explanations. We will argue that these types of explanations should be sought in the production of explainable neural models. In section 6. we summarize these alternate types of explanation, which leads us to the key claim of this work: that non-causal explanations are the *only* types of explanation that can be derived from neural models where surgical intervention fails. We walk through the implications of these findings, concluding that NLP researchers developing neural models *must* template the success conditions for explainability on the types of explanation that philosophers have identified as non-causal. We end the paper by pointing the NLP community in the direction of explanatory accounts that fit these conditions.

2. Explainable Machine Learning

In machine learning, it is generally accepted that rule-based systems are easier to interpret by both amateurs and experts compared to linear models (Lakkaraju et al., 2017); that linear models are in turn easier to interpret relative to generalized additive models (Lou et al., 2012) or Bayesian networks (Lacave and Díez, 2002); and so on until reaching the almost entirely black-box behavior of deep neural models (Miller, 2018). There is room for translation, for instance by extracting simpler proxies that can mostly replicate more complex model behavior with simple rules (Han et al., 2014); but in general, explanation in machine learning has only gotten harder over the last more than forty years (Biran and Cotton, 2017). While this hierarchy is rudimentary and fails to account for all the various dimensions of model interpretability (Lipton, 2016), it is broadly perceived to be accurate in practice. Nevertheless, a lack of explainability is rarely a barrier to implementation and widespread use. Neural models' performance continues to outpace other approaches to machine learning, and where they fall short in explanation, they make up for in performance.

But just because explanation of neural models has been difficult has not stopped researchers from trying. Much work has tried to grasp the structure of a network and what aspects of language are encoded where (Tenney et al., 2019; Clark et al., 2019). Others aim to measure how predictions change incrementally with new added information, evaluating the impact of each particular new input token in a text (Li et al., 2016). Additionally, in human-computer interaction researchers have worked to determine what users *want* from explanations (Lim and Dey, 2009), for instance by showing uses only a subset of text highlighted as important (as a simplifying step) (Bastings et al., 2019). This use of rationales, also known as *attributions*, can also be used directly at training time to encourage models to focus on or selectively ignore particular subsections of text (Dixon et al., 2018; Liu and Avci, 2019). This approach can be used without supervised span annotations, instead observing how a model responds to word deletion (Woods et al., 2017; Nguyen, 2018), and the results can be visualized using heatmaps that perform highlighting or live editing (Liu et al., 2018); generated plaintext, partially or entirely independent of the actual classification or factual content but facially plausible (Xu et al., 2015b; Liu et al., 2019); or direct exposure of structure in the underlying model, such as traversals through a graph (Yang et al., 2018; Moon et al., 2019).

In recent years, much of the hope for explanation has been pinned on *attention mechanisms*. This innovation, first introduced by Bahdanau et al. (2015), allows neural models to be trained to automatically focus on small portions of inputs, like individual sentences or even words, while making predictions. This focusing allows neural models to outperform the state-of-the-art (Yang et al., 2016) and has led to sophisticated modern architectures like the Transformer, which has currently produced the most accurate models on a wide range of tasks (Vaswani et al., 2017; Dai et al., 2019). In addition to performance gains, these layers appear to be providing human-interpretable explanations of model behavior "for free." Because the model was being trained to focus on specific subsets of information at inference time, the logic goes, it was reasonable to assume those dimensions are "most important" for the rationale of the resulting output. This approach resulted in a variety of visualizations and other attempts at model explanation being explored based, in part or in whole, on attention weights (Xu et al., 2015a; Mullenbach et al., 2018; Yang et al., 2019).

2.1. Highlighted Studies

The past year has seen skepticism emerge about this indirect, downstream use of attention. The layer was designed to facilitate increased accuracy of models — in fact, the original paper makes no claims of its human interpretability — but the field saw wide proliferation of attention's use for explanatory purposes. In this section we briefly describe the two parallel studies, and one response paper, that serve as a foundation for our analysis.

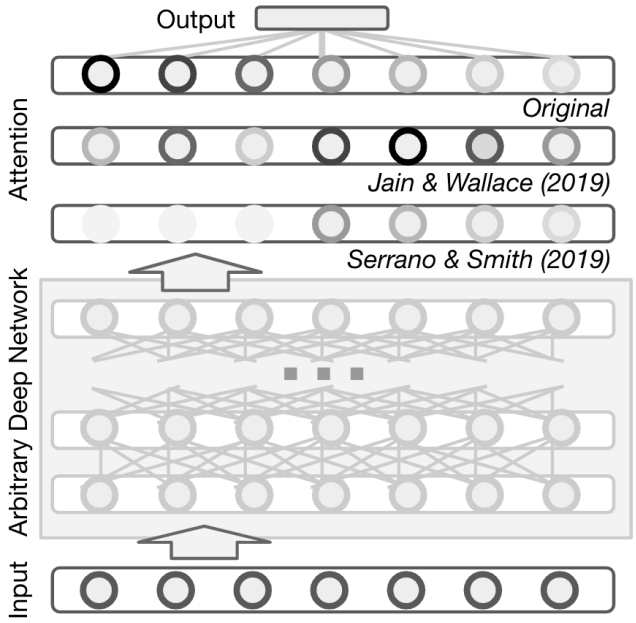

Figure 1: Researchers often use attention weights (top attention layer) to generate explanations. Jain & Wallace (middle) scramble weights and show that output remains stable; a similar result is obtained by Serrano & Smith (bottom) omitting highly-weighted nodes entirely.

"Attention is not Explanation" is an application of adversarial learning to the problem of explanation in machine learning systems. Jain and Wallace (2019) take issue with the widespread direct extraction of attention weights into visualization tools like heatmaps. They show that *counterfactual* attention weights can be discovered for a given, trained neural network. First as a proof-of-concept, the authors show that attention weights can, in some cases, be randomly scrambled without loss of performance, suggesting that "explanations" derived from those weights have little meaning. They then demonstrate an optimization problem that moves attention weights *as far as possible* away from the original attention weights of a model, without changing the model's output behavior. The authors' critique of the use of attention for explanation describes these counterfactual configurations as equally plausible, from a modeling perspective, compared to other configurations which present far more intuitive explanations. Further, they assert a strong conclusion: because there exists the possibility that these adversarial configurations can be created without changing the outputs for given inputs, we *cannot* rely upon attention as a means of explanation.

"Is Attention Interpretable?", written independently and contemporaneously, makes similar claims. Here, Serrano and Smith (2019) test attention mechanisms by manipulating the layer's weights, and show that these weights do not impact output of the model. Rather than alter weights adversarially, the authors *omit nodes entirely*. They show that a surprising number of attention weights can be zeroed out (in some cases, more than 90% of nodes), without impacting performance of the model itself. The authors test this approach with a variety of ranking methods, from random choice to sophisticated sorting based on the gradient of nodes with respect to the classifier's decision boundary. Though their results show sensitivity to these different approaches, the core finding remains: Neural models are *highly* robust to change at the supposedly crucial attention layer, producing identical outputs in a large fraction of cases even after significant alterations to the model.

But the picture is not simple. In *"Attention is not not Explanation"* Wiegreffe and Pinter (2019) produce several empirical results limiting the scope of the claims from the first two papers. The first result of that work shows that some of the classification tasks in the initial work are simply too easy for attention to matter — eliminating the layer by setting all values uniformly does not result in loss of performance. This suggests one practical boundary for attention by explanation, for any task where the additional complexity of an attention layer is wholly unnecessary to achieve state-of-the-art performance. For those tasks where attention is valuable for performance, the authors show that part of the vulnerability of manipulation to the attention layer comes from holding the rest of the model fixed. Attention as explanation, they argue, only makes sense in the context of a model that has jointly trained inner representation layers and the final attention layer. By constraining the adversary from Jain & Wallace to model-consistent behavior, they show that the resulting attention weights have much less room for modification without resulting in changes to the output.

As this debate goes on with additional empirical results, we find that computer science researchers are hampered by a lack of shared vocabulary and lack of a theoretical basis for success criteria of explanation (Lipton, 2016; Corbett-Davies and Goel, 2018). In the remainder of this paper, we advance the discussion by leaning explicitly on philosophy of science to build a more rigorous vocabulary and re-evaluate these results.

3. Philosophy and Theories of Explanation

Philosophy of science research identifies and analyzes the conceptual and logical foundations of scientific reasoning using methodology including conceptual analysis, simulation modeling, formal methods, ethnography, and case studies on historical and contemporary instances of scientific research. The nature of scientific explanation has long been a topic of central concern in philosophy of science. Philosophical research on explanation seeks to identify what explanations are — whether they are instantiated patterns of logic inference, generators of a psychological sensation of understanding, ways of encoding similar patterns of information observed in disparate systems, traces of causes and effects, or something else entirely. Along with research on laws of nature, the structure of scientific theories, the aims of science, and the role of causation in the sciences, research on scientific explanation is one of the central subdisciplines within the philosophy of science.

The aim of philosophy is not consensus-building, so a wide variety of philosophical theories of explanation continue to coexist. Some are in direct competition with one another, while others serve as complements or limiting cases of one another. This brief review, summarized in Table 1, highlights a few of the most common sorts of theories of explanation, with emphasis on the varieties that are most central

Table 1: Philosophical Theories of Explanation

	Theory	**Explananda** *(things to be explained)*	**Explanantia** *(things doing the explaining)*
Logical	Deductive-Nomological	Observed phenomenon or pattern of phenomena	Laws of nature, empirical observations, and deductive syllogistic pattern of reasoning
	Unification	Observed phenomenon or pattern of phenomena	Logical argument class
Causal	Transmission	Observed output of causal process	Observed or inferred trace of causal process
	Interventionist	Variables representing output of causal process	Variables representing input of causal process and invariant pattern of counterfactual dependence between variables
Functional	Pragmatic	Answers to why-questions	True propositions defined by their relevance relation to the explanandum they explain and the contrast class against which the demand for explanation is made
	Psychological	Observed phenomenon or pattern of phenomena	True propositions defined by their relation to the user's knowledge base and to the explanandum

to explainability in neural models. A more comprehensive overview is available in Woodward (2017).

Generally, theories of explanation may be understood as either *logical*, *causal*, or *functional*. Logical theories aim to characterize the logical structure of a cogent scientific explanation and typically emphasize the relations between explanation, laws of nature or scientific theory, and specific empirical observations. Causal theories aim to characterize explanation as an accounting of observed or expected patterns of cause and effect and are often accompanied by philosophical theories of causation itself. Functional theories, which typically focus on either the psychological or pragmatic functions of explanation, aim to characterize explanation in virtue of the function it accomplishes in scientific reasoning, rather than identifying the logical or causal structure of an explanation.

Some basic tenets of canonical theories of explanation are summarized below. Standard philosophical vocabulary for the parts of an explanation are employed: the *explanandum*, pl. *explananda*, is the thing to be explained, i.e. the target or object of an explanation; the *explanans*, pl. *explanantia*, is the thing doing the explaining.

- **Deductive-Nomological Theories** (Hempel and Oppenheim, 1948; Braithwaite, 1953; Popper, 1959; Hempel, 1967; Railton, 1978), one of the oldest logical theories of explanation, hold that explanations are deductive syllogisms. The explanantia are the premises of the syllogism, and the explanandum is the conclusion. Among the explanantia, laws of nature are always taken as the major premise, and specific empirical conditions as the minor premise.

- **Unification Theories** (Friedman, 1974; Kitcher, 1981), another logical theory, hold that explanations are not syllogistic; instead, they inhabit a more finely-structured logical space in which disparate phenomena exhibit the similar patterns of behavior. In this account, explanation consists in identifying the classification of a given argument pattern from among the accepted patterns of argument. The argument class is an explanans. Classes typically align with systems of natural laws.

- **Transmission Causal Theories** (Salmon, 1984; Dowe, 1992) characterize explanantia not as logical structures but as causal processes. These processes generate a product, which is the explanandum. Distinguishing genuine from merely apparent causal processes is a central concern of these theories and is accomplished by tracking the transmission of a signal, impulse, or mark over the course of the explanation.

- **Interventionist Theories** of *causation* introduce graph theory to the representation of causal relations and emphasize the identification of *invariance* relations between causes and effects as the target of causal claims (Woodward, 1994; Pearl, 1995; Spirtes et al., 1983; Pearl, 2000)[1]. Applied to *explanation*, the interventionist account (Woodward, 1997; Woodward, 2000; Woodward, 2001; Woodward, 2005) produces theories of causal explanation in which explananda and explanantia are connected via patterns of counterfactual causal dependence, which indicate invariant relations between purported causes and effects. This theory will be our focus in the next section.

- **The Pragmatic Theory** (Van Fraassen, 1977; Van Fraassen, 1980) contrasts itself with logical theories by characterizing explanation not as generation of a particular logical argument structure, and with transmission theories by not requiring the relay of a causal mark. Instead, the theory defines explanation functionally as answering a why-question about a phenomenon. Explanantia consist of meta-level logical structures that index explananda to an explanatory context and define relevance relations to contrasting phenomena.

- **Psychological Theories** (Trout, 2002; De Regt, 2009; Khalifa, 2012) and their critics investigate explanation not as the satisfaction of any particular argumentative structure, but rather as acts or pieces of information

[1] An additional motivation of interventionist approaches, and a challenge to causal theories of explanation, is the problem of modeling statistical patterns of dependence. For the sake of space, we do not discuss this issue further here.

that generate a sense of understanding in the agents (real or ideal) who interact with them. Significant attention is then given to characterizing what constitutes understanding. Like some of the work on explanation in neural models (Miller et al., 2017), this approach has significant overlap with the social sciences including psychology and behavioral science.

The classification system we present here is meant to capture commonly-acknowledged divisions within philosophical research on explanation. Each of the theories identified above has benefits and drawbacks, and some are more appropriate than others for capturing the sort of explanation sought in the construction of explainable neural models. This paper was initially motivated by the observation that a significant source of confusion and potential for error in constructing explainable neural models arises from failing to clarify what sort of explanation is being generated by the AI. For instance, an explanation for a model's output designed as causal fails if it only generates non-causal, nomological explanations.

4. The Interventionist Account

For philosophers, an *account* is an application of a philosophical theory to a scientific process, making explicit the set of assumptions and worldviews that are embedded in the actions, writing, or conclusions of the scientists under scrutiny. In computer science it is often the case that causal theories of explanation are preferenced. The mantra that *"correlation is not causation"* looms over scientific inquiry, description, and discussion of neural model behaviors. For this reason, deep learning explanations have narrowly focused on causal explanation. To study this, we apply the *interventionist account* of Woodward (2005).

This account focuses on those phenomena which can be explained in terms of the relationship between particular outcomes and the factors which gave rise to those outcomes. An explanation in this account relies on establishing the existence of *manipulability through intervention*. Some key features distinguish this account from others, such as logical explanations. First, the relationships between circumstances and outcomes are *empirical*, subject to data-driven verification through manipulation of those circumstances and collection of evidence. Next, that evidence is evaluated for *causality* - the dependence of the outcomes on additional variables is not merely conceptual but a direct relationship. The challenge here in explanation lies in concretely determining the existence of such a causal relationship.

In order to do so, the relationship between relevant variables in a system must be subjected to *manipulation*, where the values of those variables are changed. The theory thus lends itself well to explanations of systems which have quantifiable components, such that the quantity or value of the component can be easily denoted and modified as a variable. Performing a quantitative manipulation on a variable and then observing the changes in the output of the system as a whole, recording the overall changes through observed cause and effect, is called an *intervention*. If manipulation of quantifiable components leads to similarly quantifiable changes in output, researchers have established the first necessary, though not sufficient, elements of causal explanation.

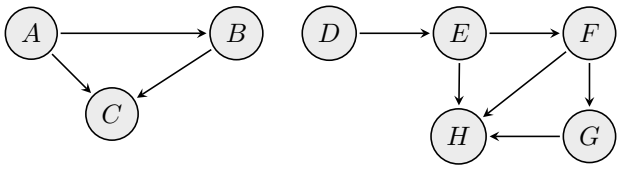

Figure 2: Network diagrams of causal systems. The system on the right resists surgical intervention between *D* and *H*.

The last piece of a successful causal explanation for a system requires that an intervention on system components is *surgical*. To define this, philosophers lean on one final concept: *invariance*. In a multivariate system, it is frequently the case that a single effect has multiple causes. In Figure 2, the diagram on the left demonstrates a simple toy system with three variables: variable A is a causal factor for both B and C. B is also a causal factor in C. To perform a *surgical* intervention explaining the relationship between A and C, holding B invariant is necessary. But the system on the right, with only a handful of variables and relationships, demonstrates that some cases *resist* surgical intervention. The only path from D to H, for instance, is indirect, passing through other causal factors. We *cannot* hold those variables all invariant while intervening on D and still cause a change in H. According to Woodward, a *surgical* intervention is an intervention that makes strategic use of invariance; a successful explanation, finally, is only one that is generated empirically through the use of surgical interventions. The relationship between D and H cannot be explained through surgical intervention.

The interventionist account is fundamentally *modal*: it relies upon counterfactuals in order to work:

> *"...an explanation ought to be such that it can be used to answer what I call a* what-if-things-had-been-different *question: the explanation must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways."* (Woodward, 2005)

What the outcome of a manipulation would be, were it to occur, is what matters for a successful explanation: a pattern of counterfactual dependence between elements of the system of variables and the output of the system. In cases where we cannot track the pattern of counterfactual dependence among variables, no amount of manipulation is sufficient to successfully explain behavior. This has clear implications for neural models, which can have hundreds of millions of parameters, interdependent in complex ways.

The interventionist account is a good fit for probing the boundaries of causal explanation in machine learning, where inputs and outputs are quantifiable as vectors, tensor elements in neural models, or probability distributions in Bayesian models. Indeed, this definition of a successful causal explanation will be familiar to researchers with experience in Bayesian statistical machine learning. Woodward's philosophical account was entwined with the development of Bayesian networks under his colleague Judea Pearl (Geiger et al., 1990), and their shared research agenda led to mathematical definitions like d-separation of variables in machine

learning. But while Pearl-style approaches to causality have been applied extensively in Bayesian models, their application remains daunting in the context of deep neural models. Overall, the interventionist account has a great deal of appeal for computer scientists. Causal relationships are intuitive and aligns to how humans learn to interact with the world; if successful explanation requires human understanding on some level, there is great potential in an account that leverages natural inclinations to modify and test existing systems. But as we shall see, causality cannot always provide adequate explanatory power for large and complex systems.

5. Applying the Interventionist Account

In what follows we will recast the findings from adversarial attention experiments from our highlighted papers in terms of Woodward's interventionist account, using the research questions introduced at the beginning of this paper.

5.1. Does the Account Apply?

Woodward's interventionist framework may be a useful way to analyze attention mechanisms in NLP; but not every philosophical theory is an appropriate fit for every scientific experiment. Before proceeding we confirm that the problem fits the conditions of a manipulation-based approach. In this case, we are looking for experiments that (1) produce empirical data, (2) hinge on causality as the core of their explanatory argument, and (3) rely on reasoning via counterfactual dependence to make causal claims.

In fact, adversarial attention configurations fit Woodward's description of intervention well. Woodward calls for modifying targeted variables in order to observe the changes to the output of the whole system, and adjustments to weights in the attention layer attempt just that: precisely shifting the focus of the algorithm to a particular segment of the input data in order to cause changes in the output, with the intent of measuring a causal effect (the outputs changing). When the system of variables represented by the attention configuration is manipulated, if there is a causal relation, the prediction generated by the model should change.

Jain & Wallace show that a vastly different attention layer which results in the same output can be found by either searching through weights for nodes, or even by scrambling the weights of the network at random. The work from Serrano & Smith is similar. Here, rather than reweighting to create an adversarial layer, an enormous number of nodes in the attention layer can be zeroed out entirely, functionally removing them from the network. Again, they test whether outputs of the model differ based on this process. Both experiments evaluate the quantitative outputs of their models on a large corpus of data, meeting the requirement for empirical evidence as part of the explanation. Both also reason about counterfactuals to develop causal claims: the development of adversarial configurations of attention weights, or adversarially zeroed-out attention nodes, both are directly evaluated for explanatory value *compared to* the attention layer that was actually learned from data (the *base*, to use Wiegreffe & Pinter's terminology).

> *A1. Yes, these studies are attempting to make arguments that fit the interventionist account.*

5.2. Is Attention Manipulation Surgical?

Jain & Wallace's central proposition is clear from their title: attention is not explanation. They make a causal argument, discovering an adversarial attention configuration which produces the same effect, resulting in a loss of uniqueness of explanation. Furthermore, the resulting adversarial weights contradict intuitions about the sources of a model's judgment. Similarly, Serrano & Smith ask whether attention is interpretable. By showing that highly-ranked attention weights can be zeroed out without affecting model performance, they argue that the answer is no. These processes initially appear to be surgical intervention: researchers assert that the initial configuration is a plausible explanation, and after manipulation, the new configuration is implausible. Two major philosophical problems appear here.

First, it is possible that there actually is a true causal relationship in *both* the adversarial and non-adversarial attention configurations, and the model predictions. In this case, the original scientists would be right to conclude that attention is not explanation: a surprising, counterintuitive causal link between two variables that should not be linked is perplexing to users. In the colloquial sense, it explains nothing. On the other hand, an explanation of a causal relationship is not necessarily *non-explanatory* just because it is *unintuitive*. If manipulations reveal the existence of adversarial attention configurations that nevertheless produce accurate predictions, it may be the result of another causal relationship between inputs and the output class being predicted. If such a causal relationship does exist but was not discovered through the model's original training, then yes, the counterintuitive attention weights raise additional challenges and questions for researchers, who must then determine how and why these two variables are linked in this way. But this does not mean the adversarial explanation is *wrong*.

To get at the deeper problem, the interventionist account offers a second and more problematic observation on the experiments. In both original studies, attention is only one part of a larger system of variables; in fact, it is the final layer, receiving as input the result of a complex series of calculations on the initial inputs. But *invariance* in all non-target variables is what makes manipulations qualify as surgical. Both highlighted papers show an unsteady relationship between input tokens and the corresponding attention weights; the relevant variables for targets of manipulation lie outside of the attention layer. The relevant system in this case is not attention alone, but attention in addition to and in connection with the neural model's prior layers. If the generation of adversarial attention configurations is possible, the interventionist account argues, then there is more at work than attention in the learned model. This is quite a big problem to overcome, as the scope of the changes to network output of the network may not match the scope of attempted interventions. Additionally, engaging in interventions on selected weights does not result in continuous, smooth changes to model outputs, especially in discrete classification tasks.

Only some manipulations are surgical, and these example studies do not meet that standard; only some sets of variables are held invariant. Wiegreffe & Pinter make this argument implicitly in their response paper, arguing that severing the attention layer from the broader training of the overall model

renders the experiment less meaningful; they argue that similar interventions on the remainder of the system are only possible with joint training between the attention layer and the rest of the model. Their critique is appropriate, and can be strengthened with philosophical vocabulary. Surgical interventions *require* explanations that depend on variables that must be held constant. Woodward's conditions for successful explanation cannot be met.

Failing this requirement has major consequences. The identification of causal relationships allows researchers to infer important details about the nature of the system which can serve in an explanation. But as the interventionist account cannot be meaningfully applied to systems which resist surgical interventions, causal explanations of the type that Woodward describes are not possible. Consequently, we will agree that attention is not explanation, but for reasons apart from, and broader than, the intuition-based arguments from Jain & Wallace. Instead we must argue that, *by definition*, attention is not explanation. The manipulation of attention manipulations cannot meet preconditions laid out as part of the boundaries of successful causal explanation.

> *A2. No, manipulating attention weights fails to meet the conditions of surgical intervention.*

5.3. Consequences of failed causal explanation

So we cannot trace the causal chain through a neural model at the level of complexity in modern NLP. Woodward's framework demands the establishment of a pattern of counterfactual dependence through the elements of a system, and this can only be demonstrated through the use of surgical interventions while tracking changes in output. Without surgical intervention, we cannot determine if a pattern of counterfactual dependence exists in the first place, or if it does, how it is constituted. Two options present themselves:

- For some reason, Woodward's causal theory is inapplicable to attention-based manipulations, and the approach is not causally problematic.
- Woodward is correct, and the absence of the possibility for surgical intervention on attention mechanisms means that a causal link between attention and model output cannot be established.

In the first case, NLP researchers are faced with a difficult question: what distinguishes our circumstances from other quantitative systems of interacting variables in science, which are adequately explained by the interventionist account? But if we choose the latter, a more practical problem emerges: In order to be explainable, an algorithm must be manipulable via surgical intervention. The results of these papers suggest a failure in principle of modern NLP networks to allow for the testing of counterfactual manipulations. This categorically renders judgment that attention-based causal explanations are destined to fail on neural models. We do not have the ability to engage in surgical intervention on attention systems at all, and consequently cannot determine the nature of the causal relationships between attention layers and the output of the system. Though we cannot determine these relationships, we *can* still conclude that attention is not explanation — but only due to the broader claim that without access to and the ability to intentionally manipulate (or hold constant) all relevant system variables, explanation is ruled out entirely.

> *A3. Attention weights alone cannot be used as causal explanations for model behavior.*

A constant in computer science, from calculating π with greater precision to mathematically complex but deterministic tasks like cryptography, has been that programs are constrained by the logic of their code, reliant on underlying notions of cause and effect. Deep learning cannot generate these causal explanations. But methods based in attention mechanisms *will* generate apparently causal explanations even where such reasoning is not possible.

Philosophical research has a grounding for these types of explanations: the *psychological* account. The success criteria for such explanations is not grounded in explanantia based in cause and effect, but in whether they produce a sense of understanding in the researcher or user of a system. The apparently causal nature of these explanations is in fact a hindrance to scientific understanding. In the context of manipulation where surgical intervention is not possible and psychological accounts take priority, the apparent causal stories are not reliable; they are causal fake news.

This undermines the central goals of explainable machine learning: to provide justification of why and how an algorithm made a decision, to hold the algorithm and its developers accountable for decisions that violate laws, and to give the subjects of those decisions actionable steps to alter the decision that the algorithm has made. If causal explanation is based in a failed methodology, all of the protections of explainability are suspect. Laws to protect members of marginalized classes will be enforced based on false understanding of model behavior; users seeking recompense will work in vain to alter their outcomes based on factors that will not produce change; and developers will allocate resources wastefully to improve model performance based on a misguided understanding of model behaviors.

But not all philosophical theories require *true* explanantia. While many theories do expect true explanantia and a testable, robust connection between explanantia and explananda, the door is opened for false but psychologically satisfying explanations. A strand of research in NLP has explicitly aimed not to generate *true* explanantia, but instead to produce *false* but *cognizable* explanantia: natural language generation, especially work using sequence-to-sequence modeling (Ehsan et al., 2018; Liu et al., 2019). This direction of research would benefit from deeper vocabulary on the relation between truth and reasoning; as is, these explanations have no theoretical grounding of functional, logical, *or* causal theories. The challenge for such research will be to articulate the conditions for success of apparently causal explanations that are known to be false.

6. Toward Non-Causal Explanation

As models with interdependent relationships among a large number of variables grow, it becomes less likely that surgical intervention on variables can be performed. Moreover, even if such such surgical interventions were still possible in principle within the model, Wiegreffe & Pinter offer compelling

concerns about the connection between attention layers and the broader model during training. To put it bluntly: the 'deep' structure of contemporary NLP is exactly what prevents causal explanation from manipulation of their parts.
Nevertheless, the user affordances attached to many explanations employ causal vocabulary, despite such research being limited to purely psychological accounts of success. If meeting the success criteria for generating an explanation means producing human-cognizable systems of causal relations between variables, the point at which explanation becomes impossible is co-extensive with the point at which the number of variables in a causal chain exceed the maximum number of relations between variables which can, in principle, be tracked by a human to whom an explanation is directed. We argue that while researchers have defined their explanations informally using the constraints and success conditions of causal explanation, they are evaluating their success instead on non-causal theories of explanation, particularly either pragmatic or psychological bases.
The practical recommendation for NLP researchers is to disentangle explainability from cause-tracking. Rather than generating causally faulty (but psychologically satisfying) explanations that pattern themselves after causal explanation, as researchers have done in the past (Ehsan et al., 2018), we urge technical researchers of explanation to pattern their notions of explanation on *non-causal* accounts. In section 3., we briefly identified non-causal accounts like the *logical* and *functional* types. These canonical philosophical theories of explanation should be part of explanation researchers' basic vocabularies. Additional, even more promising varieties of non-causal explanation come from contemporary philosophical research on explanation in mathematics and physics. Below, we summarize these accounts, derived from research at the interface between philosophy of explanation and philosophy of scientific models, and the more general research area in *optimality theories of explanation* (Strevens, 2008; Rice, 2015; Potochnik, 2018). These accounts of explanation are robust, under active study by philosophers, and are still available to NLP research:

- **Mathematical Explanations** (Pincock, 2007; Lange, 2013), iterate on logical theories of explanation. Geometric explanations use mathematical principles as the explanantia, which are taken to be modally stronger than mere causal principles or even natural laws of physics. A classic example of this type of explanation is the use of graph representations of the Bridges of Königsberg as the explanans for one's inability to cross all of the bridges exactly once in succession.

- **Structural Model Explanations** (Bokulich, 2011; Bokulich, 2018) identify scientific or mathematical models of systems as explanantia of the phenomena they represent. Explanations are built by connecting models to phenomena via a "justificatory step," whose details will be particular to the case at hand. This is a useful alternative framework for thinking about how explainable neural models will connect to the phenomena they aim to model. In early work, Sullivan (2019) has begun evaluating the current prospects for deriving understanding from machine learning.

- **Minimal-Model Explanations** (Batterman, 2001; Batterman and Rice, 2014; Rice, 2015; Rice, 2017; Rice, 2018), drawing from work on the renormalization group in applied mathematics (Wilson, 1971), generates a framework for justifying an explanation by using mathematical details to illuminate why differences between systems modeled via the same mathematics are irrelevant. By focusing on explaining away irrelevance, rather than articulating a relevance relation, these accounts flip the script for justification of purported explanations and produce a new theory of explanation of the functional sort.[2] Due to its explicit engagement with explanations whose mathematics do not map cleanly onto represented features of the system being modeled, this approach may be especially promising for NLP.

So a variety of theories of explanation remain available. However, when causal reasoning is taken off the table, some of the existing constraints from a causal conception are also placed at risk. Any non-interpretable neural network defies the sort of individuation of explanantia into a set of humanly-cognizable statements, premises, or causes, which is required for most of these theories. Because deep learning models defy this sort of individuation, recognizing which accounts of explanation work within deep learning will clarify evaluation criteria for explanation moving forward.

6.1. Limitations

While our study of explanation is based in philosophy, our contributions are based in epistemology, not in ethics. Many adjacent subfields of philosophy of science exist and only occasional interactions between computer scientists and philosophers have taken place to date (Miller, 2018; Zerilli et al., 2018). In this work, we have examined whether researchers or users are being given a true explanation. But a good *explanation* does not mean that an algorithm has made a good *decision*. Explainability research frequently studies tasks with high stakes, including notoriously biased tasks like recidivism prediction, financial risk modeling, and facial recognition for surveillance. Our work does not absolve researchers from a broader social responsibility: the presence of a successful explanation will not help if a loan is denied because of race (Fuster et al., 2018), if an accused criminal is wrongly identified because of their gender presentation (Buolamwini and Gebru, 2018), or if algorithms persecute ethnic groups (Wang et al., 2016) or misdiagnose mental health (Bennett and Keyes, 2019).
To build algorithmic decision-making in a truly socially responsible way, our work must be a component piece, incorporated into a broader foundation that accounts not only for explanation but also for ethical software development. This work provides a vocabulary for computer scientists struggling to unify the informal language that proliferates across research today, and will allow NLP researchers to improve the quality and rigor of their explanations, and provide a sure footing for the field.

[2] An alternative interpretation of renormalization-group explanations as non-causal explanations can be found in (Reutlinger, 2014). This may also prove useful to some.

References

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR*.

Bastings, J., Aziz, W., and Titov, I. (2019). Interpretable neural predictions with differentiable binary variables. In *Proceedings of ACL*.

Batterman, R. W. and Rice, C. C. (2014). Minimal model explanations. *Philosophy of Science*, 81(3):349–376.

Batterman, R. W. (2001). *The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence*. Oxford University Press.

Bennett, C. L. and Keyes, O. (2019). What is the point of fairness? disability, ai and the complexity of justice. In *Workshop on AI Fairness for People with Disabilities at ACM SIGACCESS Conference on Computers and Accessibility*.

Biran, O. and Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, page 1.

Bokulich, A. (2011). How scientific models can explain. *Synthese*, 180(1):33–45.

Bokulich, A. (2018). Searching for noncausal explanations in a sea of causes. *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*, page 141.

Braithwaite, R. (1953). *Scientific Explanation*. Cambridge University Press.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91.

Clark, K., Khandelwal, U., Levy, O., and Manning, C. D. (2019). What does bert look at? an analysis of bert's attention. In *ACL Workshop on Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP*.

Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *Synthesis of tutorial presented at ICML 2018.*

Dai, Z., Yang, Z., Yang, Y., Cohen, W. W., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of ACL*.

De Regt, H. W. (2009). The epistemic value of understanding. *Philosophy of Science*, 76(5):585–597.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73. ACM.

Dowe, P. (1992). An empiricist defence of the causal account of explanation. *International Studies in the Philosophy of Science*, 6(2):123–128.

Ehsan, U., Harrison, B., Chan, L., and Riedl, M. O. (2018). Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 81–87. ACM.

Friedman, M. (1974). Explanation and scientific understanding. *The Journal of Philosophy*, 71(1):5–19.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., and Walther, A. (2018). Predictably unequal? the effects of machine learning on credit markets. *The Effects of Machine Learning on Credit Markets (November 6, 2018)*.

Geiger, D., Verma, T., and Pearl, J. (1990). Identifying independence in bayesian networks. *Networks*, 20(5):507–534.

Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57.

Han, L., Luo, S., Yu, J., Pan, L., and Chen, S. (2014). Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. *IEEE journal of biomedical and health informatics*, 19(2):728–734.

Hempel, C. G. and Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175.

Hempel, C. G. (1967). *Aspects of Scientific Explanation And Other Essays in the Philosophy of Science*. Free Press.

Jain, S. and Wallace, B. C. (2019). Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Khalifa, K. (2012). The role of explanation in understanding. *The British Journal for the Philosophy of Science*, 64(1):161–187.

Kitcher, P. (1981). Explanatory unification. *Philosophy of science*, 48(4):507–531.

Lacave, C. and Díez, F. J. (2002). A review of explanation methods for bayesian networks. *The Knowledge Engineering Review*, 17(2):107–127.

Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. (2017). Interpretable & explorable approximations of black box models. *Proceedings of KDD Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

Lange, M. (2013). What makes a scientific explanation distinctively mathematical? *The British Journal for the Philosophy of Science*, 64(3):485–511.

Li, J., Chen, X., Hovy, E., and Jurafsky, D. (2016). Visualizing and understanding neural models in nlp. In *Proceedings of NAACL*, pages 681–691.

Lim, B. Y. and Dey, A. K. (2009). Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*, pages 195–204. ACM.

Lipton, Z. C. (2016). The mythos of model interpretability. *ICML Workshop on Human Interpretabiltiy in Machine Learning*.

Liu, F. and Avci, B. (2019). Incorporating priors with feature attribution on text classification. In *Proceedings of ACL*.

Liu, S., Li, T., Li, Z., Srikumar, V., Pascucci, V., and Bremer, P.-T. (2018). Visual interrogation of attention-based

models for natural language inference and machine comprehension. In *Proceedings of EMNLP*.

Liu, H., Yin, Q., and Wang, W. Y. (2019). Towards explainable nlp: A generative explanation framework for text classification. In *Proceedings of NAACL*.

Lou, Y., Caruana, R., and Gehrke, J. (2012). Intelligible models for classification and regression. In *Proceedings of the ACM SIGKDD*, pages 150–158. ACM.

Miller, T., Howe, P., and Sonenberg, L. (2017). Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. In *Proceedings of ICJAI Workshop on Explainable AI*.

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

Moon, S., Shah, P., Kumar, A., and Subba, R. (2019). Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of ACL*.

Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J., and Eisenstein, J. (2018). Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Nguyen, D. (2018). Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of NAACL*, pages 1069–1078.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2000). *Causality: models, reasoning and inference*. Springer.

Pincock, C. (2007). A role for mathematics in the physical sciences. *Noûs*, 41(2):253–275.

Popper, K. (1959). *The Logic of Scientific Discovery*. Hutchinson.

Potochnik, A. (2018). Eight other questions about explanation. *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*, page 57.

Railton, P. (1978). A deductive-nomological model of probabilistic explanation. *Philosophy of Science*, 45(2):206–226.

Reutlinger, A. (2014). Why is there universal macrobehavior? renormalization group explanation as noncausal explanation. *Philosophy of Science*, 81(5):1157–1170.

Rice, C. (2015). Moving beyond causes: Optimality models and scientific explanation. *Noûs*, 49(3):589–615.

Rice, C. (2017). Models donât decompose that way: A holistic view of idealized models. *The British Journal for the Philosophy of Science*, 70(1):179–208.

Rice, C. (2018). Idealized models, holistic distortions, and universality. *Synthese*, 195(6):2795–2819.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

Serrano, S. and Smith, N. A. (2019). Is attention interpretable? In *Proceedings of ACL*.

Spirtes, P., Glymour, C. N., and Scheines, R. (1983). *Causation, prediction, and search*. Springer-Verlag.

Stewart, *concurring*, P. (1964). Jacobellis v ohio. *United States Supreme Court*, 378:184.

Strevens, M. (2008). *Depth: An account of scientific explanation*. Harvard University Press.

Sullivan, E. (2019). Understanding from machine learning models. *British Journal for the Philosophy of Science*.

Tenney, I., Das, D., and Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline.

Trout, J. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science*, 69(2):212–233.

Van Fraassen, B. C. (1977). The pragmatics of explanation. *American Philosophical Quarterly*, 14(2):143–150.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, W., He, F., and Zhao, Q. (2016). Facial ethnicity classification with deep convolutional neural networks. In *Chinese Conference on Biometric Recognition*, pages 176–185. Springer.

Wiegreffe, S. and Pinter, Y. (2019). Attention is not not explanation. In *Proceedings of EMNLP*.

Wilson, K. G. (1971). Renormalization group and critical phenomena. i. renormalization group and the kadanoff scaling picture. *Physical review B*, 4(9):3174.

Woods, B., Adamson, D., Miel, S., and Mayfield, E. (2017). Formative essay feedback using predictive scoring models. In *Proceedings of the ACM SIGKDD*, pages 2071–2080. ACM.

Woodward, J. (1994). Capacities and invariance. *Philosophical Problems of the Internal and External Worlds: Essays on the Philosophy of Adolf Grunbaum*, page 283.

Woodward, J. (1997). Explanation, invariance, and intervention. *Philosophy of Science*, 64:S26–S41.

Woodward, J. (2000). Explanation and invariance in the special sciences. *The British Journal for the Philosophy of Science*, 51(2):197–254.

Woodward, J. (2001). Law and explanation in biology: Invariance is the kind of stability that matters. *Philosophy of Science*, 68(1):1–20.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

Woodward, J. (2017). Scientific explanation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2017 edition.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015a). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Xu, W., Callison-Burch, C., and Napoles, C. (2015b). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Yang, D., Halfaker, A., Kraut, R. E., and Hovy, E. H. (2016). Who did what: Editor role identification in wikipedia. In *Proceedings of the AAAI Conference on Web and Social Media (ICWSM)*, pages 446–455.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R., and Manning, C. D. (2018). Hotpotqa: A

dataset for diverse, explainable multi-hop question answering. In *Proceedings of EMNLP*, pages 2369–2380.

Yang, D., Chen, J., Yang, Z., Jurafsky, D., and Hovy, E. (2019). Letâs make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3620–3630.

Zerilli, J., Knott, A., Maclaurin, J., and Gavaghan, C. (2018). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, pages 1–23.