Causal Abstraction for Interpretable, Debiased, and Accessible Language Models

An Honors Thesis

submitted in partial fulfilment of the requirements for the degree of

Bachelor of Arts with Honours in Linguistics

presented to

the Department of Linguistics

Faculty of Linguistics

Stanford University

Chris Potts, Supervisor

by

Amir Zur

23 May 2023

This Honors Thesis represents my own work and due acknowledgement is given whenever information is derived from other sources. No part of this Honors Thesis has been or is being concurrently submitted for any other qualification at any other university.

Signed: 

Abstract

The emergent properties of increasingly large and complex language models include not only deeper language understanding, but also bias and toxicity. Recent methods in interpretability provide insight into the process by which language models arrive at a decision. In this thesis, we explore the causal effect of high-level concepts, such as a text's purpose or the gender of its subject, on a language model's output. We then consider methods for intervening on a model's abstracted causal structure in order to induce or reduce the effect of a causal variable so as to align a language model with human understanding of language. We demonstrate our method on a text-to-image scoring model, inducing a causal effect between a text's communicative purpose and the model's output for the sake of accessible image descriptions. Additionally, we explore our methods' capacities and limitations towards reducing gender bias on language models in a supervised sentiment analysis setting and an unsupervised text completion setting.

Contents

1. Introduction

In the field of computational complexity, it is a widely assumed conjecture that it is more difficult to generate a solution to a problem than to verify said solution. Consequently, it is more difficult to create artwork than appraise it, more difficult to write a book than write its criticism, more difficult to construct a proof than confirm its validity. Yet seemingly, the field of natural language processing (NLP), spurred by the development of transformer-based large language models (LLMs), is playing against these rules. LLMs generate music and images, poems and short stories, proofs and programs, that pass human judgement as artistic, creative, and factual. Yet the question of evaluating the behavior of LLMs – interpreting the generation process, detecting and assessing biases, or fact-checking statements – remains difficult and largely unanswered.

As the field of NLP continues rapidly expanding, the question of interpreting and evaluating language models is increasingly important. On the one hand, as models increase in size and complexity, their emergent abilities allow humans to rely on them for increasingly varied and important applications such as medical diagnoses, content generation/writing assistance, and K-12 education. On the other hand, as models increase in size and complexity, their emergent abilities include deeper underlying bias and toxicity, while the problem of interpreting these black-box models becomes significantly more challenging. Many researchers believe that the next key development in NLP is not to create ever larger models, but to create ever more interpretable ones [30].

Research on interpretability includes many definitions, at times conflicting, and corresponding methods. In this thesis, we focus on causal abstraction, a causal explanation method that identifies and evaluates an alignment between a high-level structural causal model and a low-level neural network. The high-level structural causal model formalizes a hypothesis about how the neural network arrives at its decision. For example, a high-level structural causal model could hypothesize that a model trained on summing three numbers would first sum the first two numbers, and then add to that sum the third number. Should a causal abstraction exist between this high-level causal model and the neural network, then the model's decision process becomes more interpretable: we can inspect and interpret

its intermediate computation, and, by intervening on it, predictably change the final sum that the model outputs. By pursuing the technique of causal abstraction on more intricate causal structures and high-level causal variables, we can hope to make progress towards interpreting the computation process of LLMs in their various applications. Hence, causal abstraction is an initial solution to the question of interpreting language models. Once we truly understand a model's behavior, though, how can we evaluate its biases and factuality?

This thesis takes the perspective that interpretability and algorithmic fairness are two sides of the same coin. The key idea is that causal abstraction provides a way to put models and humans "on the same page." That is, high-level causal models allow us to formalize human understanding of the world (e.g., common sense reasoning, contextual grounding, or ethical principles). By checking this causal model against a neural network's behavior, causal abstraction achieves a sort of knowledge sharing. For one, humans have a better understanding of a language model's behavior. Yet should the alignment exist, we can construct the right interventions on a neural network's computation process so that a language model has a better understanding of human knowledge. In this thesis, we introduce a method for manipulating the high-level causal model that a neural network implements via lightweight interventions on its intermediate computation. Our method constitutes a three-part process, similar to that of counterfactual fairness: (1) Abduction: interpret a model's behavior by finding a causal abstraction between the neural network and a hypothesized causal structure; (2) Action: intervene on the model's intermediate computation so as to modify its effect on the rest of the high-level causal graph; and (3) Prediction: let the model run to completion with the intervened intermediate output.

We demonstrate our method for interpreting and aligning a language model for two distinct applications. First, we consider image-to-text models with the goal of automatically generating accessible alt-text image descriptions. A state-of-the-art referenceless metric for image descriptions, CLIPScore [24], does not distinguish the pragmatic purpose of a text between captioning an image (i.e., incorporating context so as to supplement the image) and describing an image (i.e., providing details so as to replace the image). We induce a causal effect between the high-level purpose of a text and the outputted CLIPScore via Interchange Intervention Training (IIT) [19], a method for interpretable neural networks based on causal abstractions. Second, we consider the case of gender bias in language models. Recent research has documented underlying gender bias in state-of-the-art supervised sentiment

analysis models, as well as unsupervised text generation models [56, 38, 59, 4]. Our method makes progress towards reducing the causal effect between gender and a language model's output in certain contexts. Although our gender debiasing results face significant limitations, including treating gender as a causal variable, a narrow definition of gender, and imprecise bias evaluation, we believe that our method provides a promising direction for future research on interpretability, algorithmic fairness, and their intersection.

2. Related Work

Our work seeks to address bias in language models by connecting between interpretability and algorithmic fairness. In this section, we elaborate on these key concepts, and highlight a theoretical connection between the interpretability method of causal abstractions and the definition of counterfactual fairness.

2.1 Bias in Language Models

As with many machine learning algorithms, language models have the tendency to not only perpetuate but in fact amplify existing social biases [56, 3]. Social biases are exhibited within training data, linguistic resources, pretrained models or word embeddings, and the machine learning algorithms themselves [63, 4, 8, 16]. Concerningly, the sum of these biases is greater than its parts. Recent findings suggest that as language models grow in size and complexity, their ability to better understand and generate natural language also entails a deeper underlying conception and perpetuation of social biases [54, 5, 1, 7].

In this thesis, we explore underlying biases in state-of-the-art language models, and analyze methods for mitigating these biases. We take the perspective that a language model is biased when it exhibits a different understanding of how the world works than a human consensus [6]. In this sense, a language model might be biased with respect to real-world principles, such as by ignoring syntactic rules, physical constraints, or common sense reasoning. Likewise, a language model might be biased with respect to ethical or legal principles, such as that no decision may be based on the gender, race, age, religion, or other protected attributes of an individual [2]. Our work explores bias in grounded real-world understanding as well as ethical human understanding of language model behavior. Below we elaborate on the biases explored in our work, and related research on mitigating these biases.

2.1.1 Bias with Respect to Real-World Principles

Although LLMs display an impressive understanding of human language, these models are often limited by their ability to ground their understanding in real-world contexts [11, 31].

For example, GPT-4, the largest and most powerful language model as of time of writing, struggles with the real-world challenge of applying the order of operations to a simple expression [7]. This suggests that although GPT-4 can solve many mathematical tasks, it is biased in its approach – GPT-4 relies on "guessing" the next number in the sequence, instead of computing the intermediate outputs that a human would compute when solving the mathematical expression.

In this thesis, we explore automatically generating accessible alt-descriptions for images on the internet. As Kreiss et al. show, multimodal models are not trained to distinguish the underlying purpose of a text when scoring it against an image [33]. That is, while humans make the distinction between image *captions* (whose purpose is to supplement an image) and image *descriptions* (whose purpose is to replace that image), image-to-text models are biased towards undervaluing this distinction, and treating captions and descriptions similarly. In fact, the image-to-text scores of a state-of-the-art multimodal model, CLIP [48], do not correlate with blind or low-vision (BLV) humans, nor with sighted individuals, for the same image [32]. To address this limitation in existing multimodal models, we utilize the Concadia dataset, which consists of (image, caption, description) triplets parsed from English Wikipedia. We treat captions and descriptions drawn from the same triplet as *counterfactual* versions of each other. That is, if we were to keep everything about a caption the same (i.e., referring to the same image, used in the same context within an article), and only modified its underlying purpose from supplementing its image to replacing it, we would end up with its paired image description, and vice versa.

2.1.2 Bias with Respect to Ethical Principles

Although language models display varying social biases across many social attributes, our work focuses on evaluating and mitigating gender bias [56]. Gender biases are perpetuated within supervised learning contexts, such as sentiment analysis models [44], as well as unsupervised text generation contexts [38]. In particular, our work is concerned with gender stereotyping (category S from [10]), where models construct associations between an abstracted concept of gender and some other high-level concept (e.g., sentiment or profession) which reflect pre-existing social biases (e.g., ranking sentences with female noun phrases as being more joyful [44], or correlating the noun "doctor" with the pronoun "he" and "nurse" with "she" [38]). We view the task of debiasing as that of removing this causal connection

between a model's abstracted concept of gender and the model's final output, in contexts where the model's output should not be affected by gender.

Another way to frame the removal of a causal connection between the concept of gender and a language model's behavior is to say that when two individual inputs are identical in every way except for their gender, the model's output should remain identical. In this sense, our view of debiasing a language model is equivalent to the notion of *counterfactual fairness* [35], where counterfactual pairs differing solely across gender should be treated identically by a language model (see Section 2.3.2 for more information). In a supervised setting, the Equity Evaluation Corpus (EEC) provides counterfactual sentence pairs in the form of "he feels happy" vs. "she feels happy" [29]. In the unsupervised setting, we utilize a professions template dataset with sentence pairs such as "the actor said that" and "the actress said that", along with "the doctor said that" and "the nurse said that" [59, 4, 38]. Though we note that "doctor" and "nurse" do not constitute a counterfactual pair (they differ by more than just their gender stereotype), such pairs allow us to abstract the high-level concept of gender within a language model's computation (see Section 5 for more detail.

Lastly, we note that our work on gender bias is limited by the narrow definition of gender reflected within NLP gender bias literature. Manipulating pronouns does not fully capture the range of ways in which gender underlies natural language [56]. Furthermore, the use of a binary gender distinction does not reflect the true definition of gender, and excludes historically underrepresented transgender people [64]. We emphasize that our work is an initial step towards thinking about how to address bias in language models, and acknowledge our limitations in Section 6.

2.2 Interpretability

We take the perspective that to identify and address biases within language models, a first step is to identify and interpret the computation process of these models. While an increasingly important aspect of machine learning models, interpretability is generally an ill-defined term within machine learning literature [36]. In this section, we discuss general notions of interpretability, and expand on recent research methods for explaining the high-level computation of machine learning models.

2.2.1 Conceptions of Interpretability

Interpretability is not a monolithic term. It can imply simulatability of a model's decision-making process, transparency of a model's computation, decomposability or ability to interpret the feature space, human-in-the-loop reasoning and interaction, or *post hoc* explanations. Here we provide just a handful of examples of what interpretability could look like; consult [36] for greater breadth and analysis.

Often interpretability methods in AI are evaluated to the extent that humans can find their explanations plausible [23]. In a strict sense, such "plausability" can only arise from the ability of a human to simulate a machine learning model's computation. An interpretable model is one that "can be readily presented to the user" and understood by them [52]. While ideally useful, this notion necessarily restricts the size of an interpretable model for it to be fully simulatable and remain well-understood.

The tradeoff in model size and its inherent ability to be explained persists in the comparison between the notions of decomposability, the ability to explain each of a model's constituent parts, and transparency, the ability to understand the model's learning mechanism. While large neural networks are generally decomposable – insofar as their individual components, including their inputs, can be provided human-readable explanation [37] – their learning mechanism and optimization heuristics do not admit transparency. Meanwhile, although smaller machine learning models might allow for more human-understandable learning algorithms and interpretable model weights, such models often rely on heavily engineered feature sets that cannot be decomposed. Therefore, smaller neural networks may not necessarily be any more interpretable than large models; they are simply interpretable along a different dimension.

Simulatability, transparency, and decomposability are all important aspects for interpreting the computational mechanism of a neural network; yet, notably, human decision-making is not interpretable under any of these notions. Rather, humans often provide *post hoc* explanations to justify and reason about their decisions – why should models behave differently? *Post hoc* interpretability can consist of visualizations of significant input features [43], model-generated text explanations [39], or local explanations of model behavior. Although generally more human-accessible, *post hoc* explanations raise the concern that although a human may find such an explanation plausible and useful, there is no guarantee that plausible explanations are faithful to the model's reasoning process. Hence, the faithfulness

of an explanation to a model's underlying computation [27] is an important dimension along which to evaluate interpretability methods.

We note here that not all *post hoc* explanation methods are susceptible to violating faithfulness, nor is *post hoc* explanation distinct from transparency and decomposability. For example, feature attribution methods are a form of *post hoc* explanation method – using the model's computation graph to retrospectively explain its output – which increases the transparency of its parameters (e.g., comparison between attribution of early and later layers within a transformers-based model). In the next subsection, we will focus on this form of interpretability, and on recent methods used to better understand the computation mechanism of neural networks.

2.2.2 Interpretability of Model Computation

This line of interpretability aims to connect model activations – the intermediate steps of a machine learning model's computation graph – to human-understandable high-level features. Model computation interpretability methods include probing, gradient-based feature attribution, and causal mediation analysis, among others.

Probing methods aim to explain a black-box model by training another – often smaller – model to predict high level concepts using the black-box parameters as inputs. While often intuitive and used in practice [58, 9, 46, 26], probes are susceptible to misinterpretation. Namely, it is possible for probing methods to pick out concepts that play no causal role in the model's decision-making process [17, 15, 51]. For example, transformer-based models have rich internal representations which might capture human-understandable concepts, but may not ultimately make use of such concepts in their final output. Just because a probe picks out a meaningful human-understandable concept within a model's latent space, it does not mean that this high-level concept impacts the model's behavior.

Gradient-based feature attribution methods aim to highlight the input features with the greatest directional effect on a model's output [55, 57, 53]. Saliency mapping uses the gradient of a model's output with respect to its inputs in order to attribute the "weight" of the input's importance in the model's decision-making [55]. This approach has been made more robust through the integrated gradients algorithm, which interpolates between an input and a default "blank" input in order to account for the model's baseline computations [57]. While effective at explaining which parts of an input the model "focuses" on, feature

attribution methods do not allow for experimental "what if questions" – that is, feature attribution only provides *post hoc* explanation of low-level input features, but cannot explain the effect of high-level concepts on model behavior.

Causal explanation methods use the concepts of causality, particularly causal mediation and intervention, in order to determine a model's behavior in counterfactual scenarios [45, 59]. Such methods often utilize counterfactual inputs – i.e., inputs in which a single high-level aspect is edited while all else remains identical to the original input – in order to answer "what if" questions. For example, in order to explain a fact-checking language model, one might create a counterfactual text input in which all is kept the same except for the first word, in order to answer the question "what would the model have predicted if the first word in the input was *hello*?" A more challenging counterfactual question could be, "what would the model have predicted if the content was the same, but it was stated in a more confident manner?" Our thesis relies on this form of explanation in order to interpret and debias language models. In particular, we utilize the notion of causal abstractions in order to answer questions in the above form – and debias models so that undesired causal variables such as confidence do not play a role in a model's computation [17]. See Section 3.3 for more detail about causal abstractions.

Our perspective is that *post hoc* explanation methods that are faithful to a model's underlying computation can be used not only to understand a model's decision system, but also to intervene on that model so as to guide its high-level computation path. In the next section, we discuss definitions within the algorithmic fairness literature on guiding machine learning models to achieve fair outcomes. We hope to underscore a theoretical connection between interpretability, in the form of causal abstraction, and the notion of *counterfactual fairness*.

2.3 Algorithmic Fairness

The algorithmic fairness literature does not strive to achieve nor even define a single fairness criterion. Rather, it seeks to define constraints, definitions, and methodology that would guarantee protection from a "litany of evils" that might emerge should machine learning algorithms be left unchecked [3]. In this section, we elaborate on some of these notions and their limitations. We then focus on the notion of *counterfactual fairness* and its relation to interpretability.

2.3.1 Theories of Algorithmic Fairness

Algorithmic fairness consists of varying frameworks through which fairness can be defined and implemented. In this subsection, we discuss group fairness notions, which constitute a popular but limited guarantee of fairness; individual fairness, which is more robust but difficult to put to practice [13]; and multi-group fairness, which draws on ideas from computational complexity to design machine learning models that are "indistinguishable" from fair ones [22, 14].

Group Fairness Early notions in algorithmic fairness are known as group or statistical notions of fairness [3]. Under this framework, fairness is viewed as a constraint on model predictions, such that the model achieves a similar accuracy score across groups which differ in their sensitive attributes. For the rest of this section, let the sensitive group be S, and the remainder of the population be T (though we note that all group fairness notions below can extend to more than two groups).

One example of a group fairness notion is statistical parity, which restricts a model to assigning an equal rate of positive outcomes for the sensitive group, S, along with the rest of the population, T. In the case of admissions, this can be thought of as requiring that a model admit the same proportion of students from S as from T.

Definition 1 (Statistical Parity)**.** *Let $x \in \mathcal{X}$ represent an individual use true outcome, $y \in \{0, 1\}$, we would like to predict. A classifier $f : \mathcal{X} \to \{0, 1\}$ satisfies statistical parity if*

$$\mathbb{P}[y = 1 \mid x \in S] = \mathbb{P}[y = 1 \mid x \in T].$$

Another example, calibration, is a form of group fairness which restricts a model to "say what it means", by ensuring that its accuracy for a prediction v on a member of sensitive group S should be the same for all of the times that it predicts v for individuals outside of group S [47]. Calibration is often preferred to statistical parity in medical settings and in weather prediction.

Definition 2 (Calibration)**.** *Let $x \in \mathcal{X}$ represent an individual whose true outcome, $y \in \{0, 1\}$, we would like to predict. A classifier $f : \mathcal{X} \to \{0, 1\}$ satisfies calibration if for all $v \in [0, 1]$,*

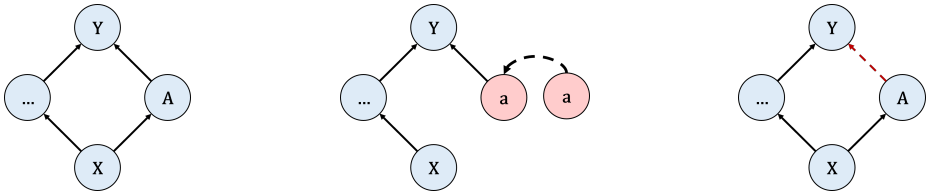
$$\mathbb{P}[y = 1 \mid f(x) = v, x \in S] = \mathbb{P}[y = 1 \mid f(x) = v, x \in T] = v.$$

Note that in calibration, the accuracy measure that is ensured across S and T is a form of precision, where we measure the rate of accuracy conditioned on the model's guess, v. Similarly, we could assert equal sensitivity and specificity, which condition on the true outcome, y; these fairness measures are denoted as equal opportunity and predictive equality, respectively, and their combination is referred to as "balance" or equalized odds [21].

Group fairness notions are useful for auditing machine learning models for bias across sensitive features. Nevertheless, they may form an incomplete and at times harmful definition of fairness. Namely, group fairness is defined as an *average* across the sensitive group and its complement. It does not allow for any distinction between the individuals within the protected group, and hence is susceptible to "stereotype threat," whereby a machine learning model learns to predict the same output for all members of a sensitive group [13]. Stereotype threat is a significant limitation of all group fairness notions, including statistical parity, calibration, and equalized odds.

Individual Fairness One notion which seeks to resolve the averaging effect of group fairness is individual fairness [13]. Under this framework, a model is fair if it "treats similar individuals similarly" (and likewise, dissimilar individuals dissimilarly). That is, the distance of a model's prediction must be proportional to the distance between its two inputs, as measured by some predefined distance metric. A significant limitation for individual fairness, hence, is the challenge of designing, learning, and computing a satisfactory metric to compute that distance between individuals. While theoretically learnable [62], such a distance metric is arguably both computationally and socially difficult to define.

Multi-Group Fairness Another algorithmic fairness framework which seeks to constrain model predictions along sensitive attributes is multi-group fairness, in particular multi-calibration [22]. Extending the notion of calibration, multi-calibration constrains a classifier to be calibrated along a rich class of attributes; this can include sensitive attributes, but also includes intersections between two classes, subsets of sensitive attribute groups (e.g., "people wearing glasses") and unions of sensitive attributes. When the group of classes is sufficiently computationally rich, the calibrated model achieves comparable levels of accuracy across any sensitive attribute and combinations thereof, and hence is not susceptible to stereotyping. In fact, provided a rich enough set of classes, a multi-calibrated predictor is computationally indistinguishable from a perfect predictor [14]. Though it is a powerful approach, multi-group fairness requires the expression of a rich set of sensitive features, which might not be easily



(a) Causal model with input variable X, sensitive attribute variable A, and output variable Y.

(b) An intervention setting the value of variable A to a.

(c) The goal of counterfactual fairness: by removing the dashed red arrow, we achieve counterfactual fairness with respect to sensitive attribute A.

Figure 2.1: Example causal model (2.1a), an intervention on this causal model (2.1b), and counterfactual fairness with respect to this causal model (2.1c).

procured or socially agreed-upon. Furthermore, learning a multi-calibrated predictor is possible, but computationally intractable. Therefore, multi-group fairness cannot be used productively in all fairness settings.

Although the algorithmic fairness frameworks above are useful within certain contexts, we do not pursue them within our work on debiasing language models. Rather, we draw a connection between work on causal interpretability methods and the algorithmic fairness notion of counterfactual fairness.

2.3.2 Counterfactual Fairness

We believe that the algorithmic fairness framework of counterfactual fairness is directly applicable to research within causal explanation and debiasing [35]. Under this framework, a model is fair if its prediction for an input x is equal to its prediction for a counterfactual edit of x in which only its sensitive attribute is altered.

Counterfactual fairness is a causal notion of algorithmic fairness, which assumes access to a causal graph that represents existing human knowledge. Following Pearl [45], a causal graph is a triple (U, V, F). The variables of the causal graph consist of U, the latent background variables, and V, the observable variables that we can measure and manipulate. The connection between variables in the causal graph is determined by a set of functions, $F = \{f_1, \ldots, f_n\}$, one for each $V_i \in V$, such that the value of each variable V_i is determined by $V_i = f_i(pa_i, U_{pa_i})$ (pa_i is the set of parents of V_i, i.e., nodes with outgoing edges towards V_i). The set of functions, F, is known as the structural equations of the causal graph. See Figure 2.1 for an example of a causal graph.

The key to counterfactual fairness is the ability to conceptualize and utilize counterfactual quantities. That is, we would like to be able to evaluate statements such as "the value of Y if A had taken on the value of a". By assumption, the state of any observable variable is fully determined by the background variables and structural equations (even if we cannot directly compute the value of the background variables). Hence, we denote the counterfactual value of Y had A taken on the value of a as $Y_{A\leftarrow a}(u)$. At times we simplify the notation to $Y_{A\leftarrow a}$.

Inferring the counterfactual $Y_{A\leftarrow a}$, given some evidence b, means computing the probabilities $\mathbb{P}[Y_{A\leftarrow a} \mid B = b]$. Inference proceeds in three steps: (1) Abduction: for a given prior on the background variables U, compute the posterior of U given the evidence $B = b$; (2) Action: intervene on A by substituting the equations for A with the value a, resulting in the modified structural equations F_a; (3) Prediction: let the causal model run to its completion in order to compute the desired probability $\mathbb{P}[Y_{A\leftarrow a} \mid B = b]$.

Now that we have established definitions for counterfactual quantities and the process by which we can estimate them, we provide the definition of counterfactual fairness. The key notion of counterfactual fairness is that, given an input $X = x$ and a sensitive attribute $A = a$, changing the value of A to a' should not change the value of the final model output, Y. Supposing that A represents gender, counterfactual fairness seeks to guarantee that individuals who are completely identical except for their gender must be treated identically by a machine learning model.

Definition 3 (Counterfactual Fairness)**.** *Let A be a sensitive attribute (e.g., gender), and let X be the remaining features which specify an individual input. Suppose that we are given a causal model (U, V, F) where $V = A \cup X$. We say that predictor $\hat{Y}$ is counterfactually fair if under any input $X = x$ and any sensitive attribute value $A = a$, we have*

$$P\left[\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a\right] = P\left[\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a\right]. \tag{2.1}$$

for all y and for any value a' attainable by A.

A key challenge to counterfactual fairness is its dependence on a structured causal model (SCM), which defines the impact of changing a sensitive attribute S on the input generation process. Oftentimes, causal models are not readily available, and require strong assumptions about natural or social processes. We believe that causal explanation methods such as those described in Section 2.2.2 can help address this challenge in counterfactual fairness. Namely,

causal explanation methods can be used to estimate the causal structure within a blackbox neural network; once this causal structure is uncovered, definitions of counterfactual fairness such as Definition 3 can be applied in order to debias the existing model.

Indeed, we take the perspective that causal interpretability and counterfactual fairness are two sides of the same coin. By being able to answer "what if" questions (i.e., compute counterfactual quantities), we have the potential to intervene on a model's abstracted computational model and ensure that it is counterfactually fair across critical sensitive attributes. In the next section, we elaborate on causal abstractions, and demonstrate that counterfactual fairness can be thought of as a special case of a complete causal abstraction.

3. Causal Abstraction

The theory of causal abstraction allows us to reason about a language model's computation as a low-level causal path, which can be abstracted by a human-interpretable high-level causal structure. Through this abstraction to a high-level causal graph, we can interpret a model's decision pathway, and intervene on its intermediate computation in order to predictably modify its behavior. Hence, through causal abstractions, we make a theoretical connection between interpretability and fairness. In this section, we demonstrate the concept of causal abstraction – and its connection to counterfactual fairness – with a running example on debiasing a fact-checking language model.

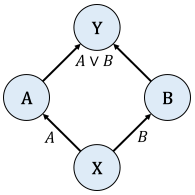
3.1 Structural Causal Models

Causal abstraction defines an alignment between two structural causal models, such that the first high-level model can abstract the second model. In this subsection, we provide formal definitions of structural causal models (SCMs), and the key intervention operation.

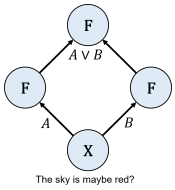
3.1.1 Structural Causal Model

A structural causal model $\mathcal{M}$ with causal variables $\mathcal{V}$ is a directed acyclic graph (DAG), as in Figure 3.1. Each causal variable $V \in \mathcal{V}$ can take a value within a set of possible values $\texttt{Val}(V)$, and has a structural equation F_V that sets the value of V based on the values of its parents, PA_V. The input to a causal graph is the set of variables with no parents, $\mathbf{V}_{in}$. Likewise, the output of a causal graph is the set of variables with no children, $\mathbf{V}_{out}$. In our work, a structural causal model $\mathcal{M} = (V, PA, \texttt{Val}, F)$ can represent both symbolic computations and neural networks.

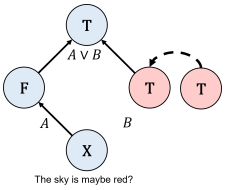
To compute the output of a causal model, we begin with input values for each of our input variables, $\textbf{input} \in \texttt{Val}(\mathbf{V}_{in})$. We then evaluate each of the structural equations F, in order from $\mathbf{V}_{in}$ to $\mathbf{V}_{out}$, and finally output the values for $\mathbf{V}_{out}$. More generally, we define $\texttt{GETVALS}(\mathcal{M}, \textbf{input}, \mathbf{V}) \in \texttt{Val}(\mathbf{V})$ to be the values that $\mathbf{V}$ take on when computing $\mathcal{M}$ on **input**.



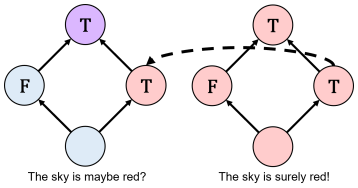
(a) Example causal model with input variable X, intermediate variables A and B, and output variable Y where $Y = A \vee B$.



(b) Example computation of the causal model in 3.1a on a nonfactual and unconfident input sentence.



(c) Example of an intervention setting $B = T$ on an input sentence where $A = F$ and $B = F$ (i.e., "what if X was the same, but stated more confidently?").



(d) Example of an interchange intervention between a counterfactual pair that differ in their confidence level, B.

Figure 3.1: Example causal model (3.1a), a computation of this model (3.1b), an intervention on this model (3.1c), and an interchange intervention (3.1d).

3.1.2 Intervention

The key operation that we can perform on a structural causal model $\mathcal{M}$ is an intervention. An intervention sets the value of a given set of causal variables $\mathbf{V}$ to some predefined set of values $\mathbf{v}$. More formally, we perform an intervention $\mathcal{M}_{\mathbf{V} \leftarrow \mathbf{v}}$ by constructing a causal model identical to $\mathcal{M}$, except that the structural equations for $\mathbf{V}$ are set to the constant values $\mathbf{v}$. This definition corresponds closely to the `do` operator of [45], which characterizes interventions on models for the purpose of exploring hypothetical or counterfactual states.

3.1.3 Example Causal Model

Suppose that we are training a language model to serve as a fact-checker. That is, given some input sentence, for example "The sky is blue", our language model decides whether the sentence is factual (in which case it outputs 1) or farcical (in which case it outputs 0). However, we are wary that our language model is biased by a sentence's confidence. That is, when a statement is stated confidently, for example "the sky is definitely red!", our model tends to output 1 regardless of whether it is a truly factual statement. In order to interpret our model's behavior, we decide to construct a structural causal model $\mathcal{M} = (V, PA, \texttt{Val}, F)$

to serve as our *hypothesis* about the underlying computation that our language model performs.

Our structural causal model consists of four causal variables: X, which represents our input sentence, and whose value can be any natural language sentence; A, which represents whether the input sentence is factual, and whose value can be T (i.e., X is factual) or F (i.e., X is farcical); B, which represents whether the input sentence is stated confidently, and whose value can likewise be T or F; and Y, which represents the output of our model, and whose value can be one of T or F. We are confident that our model will output T when the input sentence is truly factual. However, we are worried that it might also output T when the input sentence is just confidently stated. Hence, we hypothesize that our language model follows a causal structure in which the value of Y is defined by $F_Y = (A \vee B)$. That is, we hypothesize that the language model will output T if either (1) the statement is indeed factual or (2) the statement is confidently stated, or both.

Figure 3.1c demonstrates the intervention $\mathcal{M}_{B \leftarrow T}$ on our causal model. After the intervention, the final output is $\texttt{GETVALS}(\mathcal{M}, \textbf{input}, Y) = T$. Note that this intervention changed the value of our final output, meaning it *mediates* between the input causal variable and the output causal variable. Another way to state this is that, counterfactually, had the sentence "The sky is maybe red?" been stated confidently, we hypothesize that our model would predict that it is factual.

3.2 Interchange Interventions

Now that we have set up a hypothesis about the computation process of our language model in the form of a structural causal model, we have two ways to view counterfactual quantities. The first way is an input-level counterfactual, in which we construct a sentence that is identical in every sense except for the high-level feature that we care about. For example, the sentence "The sky is surely red!" is a counterfactual of the sentence "The sky is maybe red?" where the value of B, the confidence level, is counterfactually edited from $B = F$ to $B = T$. The second way is by intervening on our high-level causal graph, which generates the graph $\mathcal{M}_{B \leftarrow T}$ as in Figure 3.1c. Access to both types of counterfactual states allows us to perform an interchange intervention, which is the key step in causal abstractions.

3.2.1 Interchange Intervention

The idea behind an interchange intervention is to use the intermediate causal value computed on a **source** input in order to intervene on the computation of a **base** input in a given causal model – hence "interchanging" the intermediate computation of **base** with that of **source**. Formally, given a set of variables $\mathbf{V}$ for which we would like to compute an interchange intervention, we construct a new causal model

$$\mathcal{M}_{\mathbf{V} \leftarrow \texttt{GETVALS}(\mathcal{M}, \mathbf{source}, \mathbf{V})}.$$

Running this intervened model on the **base** input all the way to completion (i.e., computing the value of $\mathbf{V}_{out}$, we get the desired interchange intervention value. Putting these steps together, we obtain the interchange intervention

$$\texttt{INTINV}(\mathcal{M}, \mathbf{base}, \mathbf{source}, \mathbf{V}) := \texttt{GETVALS}(\mathcal{M}_{\mathbf{V} \leftarrow \texttt{GETVALS}(\mathcal{M}, \mathbf{source}, \mathbf{V})}, \mathbf{base}, \mathbf{V}_{out}).$$

In short, the interchange intervention provides the output of the model $\mathcal{M}$ for the input base, except the variables $\mathbf{V}$ are set to the values they would have if **source** were the **input**. Note that if **source** is a counterfactual of **input** where only the values of $\mathbf{V}$ were changed, then the output of $\texttt{INTINT}(\mathcal{M}, \mathbf{base}, \mathbf{source}, \mathbf{V})$ is the same as running $\mathcal{M}$ on **source**. This property of interchange intervention allows us to utilize counterfactual input in order to search for and evaluate alignments across structural causal models, as we explain in Section 3.3.

3.2.2 Example Interchange Intervention

What would an interchange intervention look like for our biased fact-checker? Suppose that we have access to a counterfactual pair of inputs, such as those mentioned earlier in this section. That is, let our **base** input be "The sky is maybe red?" and our **source** input be "The sky is surely red!" One question we might like to answer is, what do we expect our language model to predict, if the confidence level of these two sentence was to be "swapped" (that is, the **base** sentence would gain the confidence level of the **source** sentence)? Performing an interchange intervention on our high-level causal graph, as in Figure 3.1d, would give us the answer!

Additionally, we can confirm that the value of this interchange intervention must be equal to the value of computing our causal model on the **source** input on its own. This is because, since **source** is a counterfactual of **base**, everything about their computation must be equal except for the confidence level; after swapping the confidence level of **base** with that of **source**, there is nothing to tell **base** apart from **source**! In the next section, we will utilize the equivalence between interchange intervention and counterfactual behavior in order to interpret our biased language model.

3.3 Causal Abstraction

A causal abstraction occurs when we have an alignment between two causal models, a high-level model and a low-level model. The key idea behind a causal abstraction is that any computation model, including the language model that we seek to debias, can be thought of as a structural causal model. Hence, we can utilize the concept of causal abstraction to interpret whether a low-level language model implements a hypothesized high-level structural causal model.

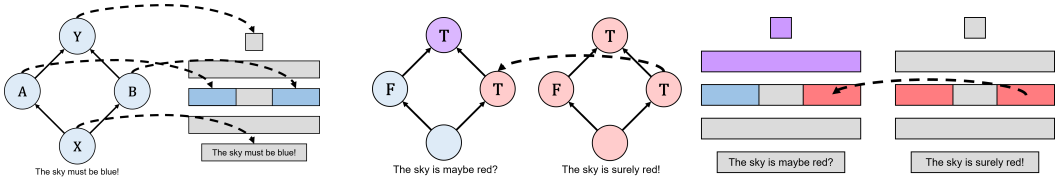
3.3.1 Causal Abstraction

Suppose that we have a high-level structural causal model $\mathcal{M}_{\mathcal{H}}$ and a low-level structural causal model $\mathcal{M}_{\mathcal{L}}$. We say that $\mathcal{M}_{\mathcal{H}}$ is a causal abstraction of $\mathcal{M}_{\mathcal{L}}$ if there exists some alignment between their intermediate states, Π, such that the behavior of $\mathcal{M}_{\mathcal{H}}$ is equivalent to the behavior of $\mathcal{M}_{\mathcal{L}}$ up to interchange interventions.

Formally, consider structural causal models $\mathcal{M}_{\mathcal{H}}$ and $\mathcal{M}_{\mathcal{L}}$ with identical input and output spaces[1]. Let an alignment Π be a mapping from intermediate variables in $\mathcal{V}_{\mathcal{H}}$ to non-overlapping subsets of variables in $\mathcal{V}_{\mathcal{L}}$. Now, consider some intermediate variable $V_H \in \mathcal{V}_{\mathcal{H}}$, and define ${\mathcal{M}_{\mathcal{H}}}^*$ to be $\mathcal{M}_{\mathcal{H}}$ with every variable marginalized other than $\mathbf{V}_{inH}, \mathbf{V}_{outH}$, and V_H. We say that the high-level model $\mathcal{M}_{\mathcal{H}}$ is a causal abstraction of the low-level model $\mathcal{M}_{\mathcal{L}}$ if, for all base and source values $\mathbf{b}, \mathbf{s} \in \mathbf{V}_{inH}$,

$$\texttt{INTINV}({\mathcal{M}_{\mathcal{H}}}^*, \mathbf{b}, \mathbf{s}, V_H) = \texttt{INTINV}(\mathcal{M}_{\mathcal{L}}, \mathbf{b}, \mathbf{s}, \Pi(V_H)). \tag{3.1}$$

[1]Note that the assumption of identical input and output spaces is not limiting, since we can always introduce extra causal variables within $\mathcal{M}_{\mathcal{H}}$ to map from the input space or into the output space of $\mathcal{M}_{\mathcal{L}}$.



(a) Example causal alignment between the high-level causal structure in Figure 3.1 and a low-level language model.

(b) Example interchange intervention between counterfactual sentences that differ in their level of confidence, performed on a high-level model and its aligned low-level model.

Figure 3.2: Example causal alignment between the high-level causal structure in Figure 3.1 and a low-level language model, and an interchange intervention performed according to this alignment.

That is, $\mathcal{M}_\mathcal{H}$ is a causal abstraction of $\mathcal{M}_\mathcal{L}$ since the behavior of $\mathcal{M}_\mathcal{L}$ can be abstractly implemented by computing the output of $\mathcal{M}_\mathcal{H}$. To see this, consider the base when $\mathbf{b}$ and $\mathbf{s}$ are counterfactual inputs with respect to V_H. The causal abstraction relationship in Equation 3.1 states that the interchange intervention value of ${\mathcal{M}_\mathcal{H}}^*$ on $\mathbf{b}, \mathbf{s}$ is identical to the interchange intervention value of $\mathcal{M}_\mathcal{L}$ on this counterfactual pair. But, as we've seen in the previous subsection, the value of this interchange intervention is equivalent to the output value of running ${\mathcal{M}_\mathcal{H}}^*$ with an intervention on V_H. Hence, Equation 3.1 entails that there is a complete correspondence between some set of low-level causal variables $\Pi(V_H)$ and the high-level causal variable V_H, such that the causal effect of $\Pi(V_H)$ on $\mathcal{M}_\mathcal{L}$ is equivalent to the causal effect of V_H on $\mathcal{M}_\mathcal{H}$. Therefore, causal abstraction captures the causal effect of a high-level, human-interpretable variable on the computation of a low-level model.

3.3.2 Example Causal Abstraction

We return to our high-level causal model, visualized in Figure 3.1, which hypothesizes that our fact-checking language model is biased towards believing any input text which is confidently stated. Let that high-level causal model be $\mathcal{M}_\mathcal{H}$, and let our biased language model be $\mathcal{M}_\mathcal{L}$. Our variable of concern here is the high-level causal variable B, which measures whether the input sentence is confidently-stated or not. What would it look like for our $\mathcal{M}_\mathcal{H}$ to be a causal abstraction of the language model $\mathcal{M}_\mathcal{L}$, with respect to our variable of concern B?

Let Π be an alignment between the variables in $\mathcal{M}_\mathcal{H}$ and the activations of $\mathcal{M}_\mathcal{L}$. For example, suppose that $\mathcal{M}_\mathcal{L}$ is a three-layer transformer with hidden dimension 36. Let Π map A to the first 12 neural activations in the second layer, and map B to the last 12 neural

activations in the second layer, as illustrated in Figure 3.2a. Now, suppose that $\mathcal{M}_{\mathcal{H}}, \mathcal{M}_{\mathcal{L}}$, and Π satisfy Equation 3.1 with respect to the counterfactual pair $\mathbf{b}$ = "The sky is maybe red?" and $\mathbf{s}$ = "The sky is surely red!". In this case, we could perform an interchange intervention between our low-level language model and high-level causal model, and expect the outputs to be identical, as illustrated in Figure 3.2b.

Should Equation 3.1 hold for all possible inputs $\mathbf{b}$ and $\mathbf{s}$, then we have a causally verified interpretation of our language model: the language model computes both the truthfulness and the confidence level of the input sentence, stored in the neural activations of its second layer, and outputs 1 if either the sentence is factual or if it is confidently stated. By intervening on the neural activations that align with the high-level concept of confidence, $\Pi(B)$, we can now answer counterfactual questions in the form of, "what would the language model output have been, had the input sentence been stated with a different degree of confidence?"

We acknowledge that in practice, it is not feasible to verify Equation 3.1 on all possible pairs of inputs. Rather, supposing access to some dataset with counterfactual pairs, one can estimate Equation 3.1 by computing the interchange intervention accuracy, or the rate at which the interchange intervention output of the low-level model is equal to the interchange intervention output of the high-level model. Interchange intervention accuracy is used to evaluate possible alignments between $\mathcal{M}_{\mathcal{H}}$ and $\mathcal{M}_{\mathcal{L}}$. In future sections, we discuss how to either (1) induce a high interchange intervention accuracy given a predetermined alignment or (2) search for an alignment which achieves a high interchange intervention accuracy.

Through causal abstraction, we have discovered a method to interpret our language model via answering counterfactual questions with respect to our causal variable of concern (i.e., the confidence level of the input sentence). However, is there a way for us to remove the causal effect of this variable completely? That is, can we find the way to answer the question, "what would the language model output have been, had it not cared about the degree of confidence in the input sentence in the first place?"

3.4 Causal Abstraction and Counterfactual Fairness

In this section, we present a theoretical connection between causal abstraction and counterfactual fairness. Our key idea is that, should a causal abstraction exist between some high-level model $\mathcal{M}_{\mathcal{H}}$ and a low-level model $\mathcal{M}_{\mathcal{L}}$ with respect to a sensitive attribute A,

then running $\mathcal{M}_{\mathcal{L}}$ with a fixed intervention value for $\Pi(A)$ results in a low-level model which achieves counterfactual fairness.

3.4.1 Counterfactual Fairness

Let us revisit the counterfactual fairness condition specified in Definition 3. Note that Definition 3 asserts that the final output of a predictor $\hat{Y}$, given some input X, remains the same no matter the intervention value of A. But now, suppose that we have access to the internal activations of predictor $\hat{Y}$, we have a causal abstraction between $\hat{Y}$ and the high-level causal variable A. More formally, let $\mathcal{M}_{\mathcal{L}} =: Y$. Let $\mathcal{M}_{\mathcal{H}}$ be a structural causal model with input X, high-level causal variable A, and output variable Y, where $F_Y(U) =: \hat{Y}(U)$ (i.e., $\mathcal{M}_{\mathcal{H}}$ mimics the behavior of the predictor $\hat{Y}$). Now, suppose that we have an alignment Π which satisfies Equation 3.1 with respect to all base and source pairs. That is, we have a complete causal abstraction of the effect of A on $\hat{Y}$.

We claim that any intervention on A in the computation pathway of $\hat{Y}$ results in a model which satisfies counterfactual fairness. Consider the intervened-upon model ${\mathcal{M}_{\mathcal{L}}}^* =: \mathcal{M}_{\mathcal{L}_{\Pi(A) \leftarrow \Pi(a)}}$ for some value a. Since we have removed the causal pathway from the input to the computation of A in ${\mathcal{M}_{\mathcal{L}}}^*$, the input value of A no longer affects the output of ${\mathcal{M}_{\mathcal{L}}}^*$. Hence, ${\mathcal{M}_{\mathcal{L}}}^*$ satisfies Definition 3 with respect to A.

Therefore, we conclude with a theoretical connection between interpretability, in the form of causal abstraction, and algorithmic fairness, in the form of counterfactual fairness. A completely causally interpretable predictor is also a counterfactually fair predictor: all that is required is a fixed intervention on the high-level abstraction of the sensitive attribute which we would like to protect.

3.4.2 Example Counterfactual Fairness

Let us return one last time to our biased, but now interpretable fact-checking language model. How could we debias our model with respect to the confidence level of its inputs?

As we've seen in the previous subsection, by intervening on $\Pi(B)$ (i.e., the range of neural activations corresponding to the high-level concept of confidence level), we can predictably influence our model's output and "override" the confidence level specified in the input. What if we were to use the same intervention, then, no matter the input? The key idea in this thesis is that, by intervening on the computation of the high-level concept of confidence so

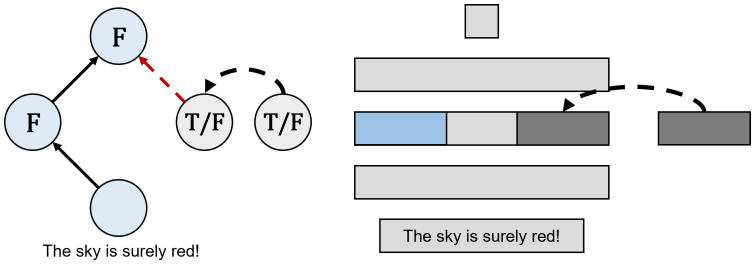


Figure 3.3: Example intervention on our low-level language model, visualized by a gray rectangle, and the aligned intervention on the high-level causal model from Figure 3.1.

that it remains exactly the same no matter the input, we can guarantee that the confidence level of the input has no causal effect on our model's output. That is, by setting a fixed intervention, we construct a low-level language model that is agnostic to its input's degree of confidence.

For our example, we can come up with a vector with 12 entries, say $b = (1, 2, ..., 12)$, and construct an updated language model $\mathcal{M}_{\mathcal{L}}{}^{*} = \mathcal{M}_{\mathcal{L}\Pi(B)\leftarrow b}$ as illustrated in Figure 3.3. In our new language model, every input is computed identically to how it would be computed by the original language model $\mathcal{M}_{\mathcal{L}}$, but the activation of the high-level concept of confidence is replaced by our fixed b vector. In our simplified diagram, we do not have a sense as to whether b represents "confidence" or "lack of confidence" (i.e., whether $\Pi(b) = T$ or $\Pi(b) = F$). However, for the sake of counterfactual fairness, the effect of b on the behavior of our model is not important, because it is independent of the input. Consider an arbitrary counterfactual pair of inputs $\mathbf{b}$ and $\mathbf{s}$, where $\mathbf{s}$ differs from $\mathbf{b}$ only in its degree of confidence. We now have a guarantee that, with our updated language model, $\mathcal{M}_{\mathcal{L}}{}^{*}(\mathbf{b}) = \mathcal{M}_{\mathcal{L}}{}^{*}(\mathbf{s})$. The value of $\mathcal{M}_{\mathcal{L}}{}^{*}(\mathbf{b})$ might not be aligned with whether $\mathbf{b}$ is truly factual, but it is no longer by whether $\mathbf{b}$ is confidently stated. Hence, we have an interpretable and fair language model, thanks to the concept of causal abstractions!

Finally, we note that although counterfactual fairness specifies that the behavior of a predictor should be identical on counterfactual input, counterfactual fairness does not specify what this behavior should be. Hence, any intervention that we choose on our high-level variable will achieve counterfactual fairness. This prompts the question, which intervention should we choose? In Section 5 we revisit this question and explore methods for designing interventions while maintaining the original capabilities of our biased language model.

We are now ready to explore methods which achieve causal abstraction and counterfactual

fairness on language models. In Section 4, we introduce a method for debiasing multi-modal models with respect to the purpose of a text. In Section 5, we introduce a method for debiasing instruction-following language models with respect to gender pronouns.

4. Inducing Causal Effect: Accessible Image Descriptions

Recently developed multi-modal models have the potential to improve internet accessibility for blind or low-vision (BLV) individuals by providing alt-descriptions for images. However, these multi-modal models are trained on data that does not distinguish the purpose of a text [33]. In particular, these models fail to distinguish between a caption that is meant to complement an image and a description that is meant to replace an image for the purpose of accessibility. See Figure 4.1 for the contrast between an example image description and image caption [33]. In this section, we address this form of bias through Interchange Intervention Training (IIT), a method for inducing causal abstraction, without comprising the capabilities of the original multi-modal models. Our analysis focuses on CLIP [48], a state-of-the-art model for computing image-to-text similarity.

4.1 Image Descriptions with a Purpose

Multi-modal models constitute a promising tool for image description generation. However, these models are not sensitive to the communicative purpose behind a piece of text, which is critical for producing accessible alt-descriptions of images [32, 33]. In this thesis, we adapt a referenceless metric, CLIP, to distinguish the purpose behind a text, which improves its

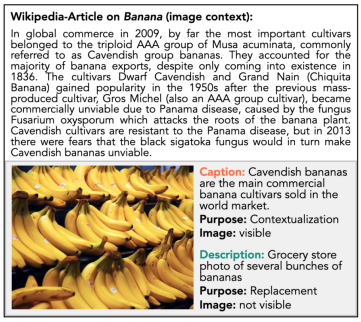


Figure 4.1: From Concadia [33], an example of an image with two associated texts: the description as provided in the image's alt text, and the caption as displayed below the image in the article. For captions, the image content is presupposed whereas descriptions aim to stand in for the image.

suitability for the evaluation of image description models.

In order to evaluate image description generation, researchers face the trade-off between using human evaluations, which can be time-consuming and difficult to obtain, and automated evaluations (e.g. CLIPScore, CIDEr) which are not sensitive to context nor the specific task at hand. Indeed, [32] evaluate the correlation between CLIPScore and human evaluations of image alt-descriptions, finding that CLIPScore neither correlates with the evaluations of sighted evaluators nor with those of BLV evaluators. Despite achieving high performance accuracy on a plethora of image-text alignment tasks, CLIP is as of yet unsuitable for assessing alt-text generation models.

One reason for the poor performance of CLIP on image description evaluation is that its training data and objective do distinguish between image descriptions and image captions, which [33] assert is critical in practice. In order to motivate this distinction, the authors release Concadia, a dataset consisting of 96,918 images with corresponding descriptions, captions, and surrounding context. In our work, we treat description-caption pairs as *counterfactual* pairs, whose entities are identical (i.e., describing the same image) except for their *communicative purpose* – captions are meant to supplement the image, whereas descriptions are meant to replace it entirely.

4.2 Interchange Intervention Training

Interchange intervention training (IIT) is a finetuning paradigm that localizes high-level causal variables within a neural network's intermediate representations in order to induce a causal abstraction relationship between the neural network and a predefined structural causal model. In our setting, we utilize IIT to induce a causal effect between the purpose behind a text (i.e., describing versus captioning) and the outputted similarity score.

4.2.1 Our Structural Causal Model

To finetune CLIP with IIT, we first need to define a structural causal model that CLIP should adhere to. Rather than provide a full structural causal model $\mathcal{M} = (V, PA, \texttt{Val}, F)$, which requires structural equations F, we provide a *partial* causal model as shown in Figure 4.2.

Formally, suppose that we have an image we wish to describe. Then our high-level causal structure $\mathcal{M}_{\mathcal{H}}$ consists of four causal variables: X, which represents our input text, and

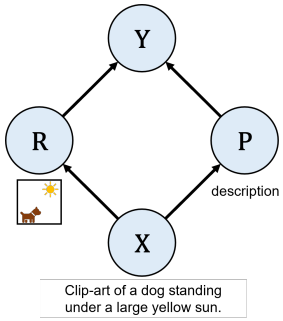


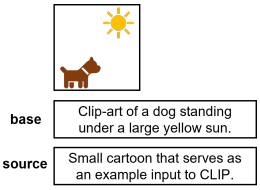
Figure 4.2: High-level causal model for the purposes of debiasing a language model with respect to the communicative purpose of a text. An example computation is provided below the high-level causal variables.

whose value can be any natural language sentence; R, which represents the relevance of the text to the image, and can take on any real value; P, which represents the communicative purpose of the text, and whose value can be one of desc. or capt.; and Y, which represents the similarity between the input text and our image. Crucially, we do not explicitly define the causal effect of the text's purpose, P, on the final similarity score, Y. Rather, we only define the direction of the causal effect: if we have a counterfactual pair of inputs $X_{P \leftarrow \text{desc.}}$ and $X_{P \leftarrow \text{capt.}}$ that only differ in their communicative purpose, then $F_Y(X_{P \leftarrow \text{desc.}}) > X_{P \leftarrow \text{capt.}}$. That is, our high-level structural causal model behaves identically to CLIP, but consistently prefers texts whose underlying purpose is to replace the image over texts whose underlying purpose it to supplement it with additional context.

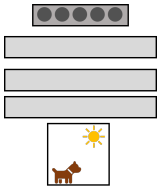
In Section 3.3, we assumed knowledge of an alignment between our high-level causal structure and the low-level model we seek to debias, such that the high-level model is a causal abstraction of the low-level model. However, since we do not expect CLIP to be abstracted by our ideal high-level causal structure (i.e., CLIP is not sensitive to the communicative purpose of a text input), we do not expect to find such an alignment. Rather, we modify interchange intervention training (IIT), which *induces* a causal abstraction relationship given a pre-determined alignment.

4.2.2 Contrastive IIT

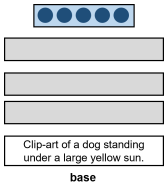
Given access to a dataset D with counterfactual pairs, a low-level neural network $\mathcal{N}^\theta$, a high-level causal model $\mathcal{M}_\mathcal{H}$, and a predetermined alignment between the two models Π, the IIT method induces a causal abstraction relationship between the structural causal models with respect to the alignment. The key idea is that IIT finetunes the neural network with



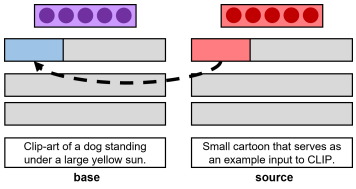
(a) Given: an image, description, and caption triplet.



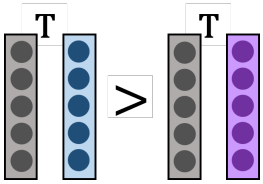
(b) First, run a forward pass of the CLIP image encoder in order to compute a static image representation.



(c) Next, run a forward pass of the CLIP text encoder with the base input.



(d) Now, perform an interchange intervention between the base and the counterfactual source text input, according to our prespecified alignment.



(e) Lastly, compute the contrastive learning objective loss as in Equation 4.2.

Figure 4.3: Example forward pass of interchange intervention training (IIT) [19], adapted to the setting of CLIP. Backpropagating through the computed CLIP objective will induce the causal structure visualized in Figure 4.2.

the interchange intervention objective

$$\sum_{\mathbf{b},\mathbf{s}\in D} \texttt{Loss}(\texttt{INTINV}(\mathcal{M}_{\mathcal{H}}, \mathbf{b}, \mathbf{s}, V), \texttt{INTINV}(\mathcal{N}^{\theta}, \mathbf{b}, \mathbf{s}, \Pi(V))). \tag{4.1}$$

which optimizes the interchange intervention accuracy with respect to some loss function, such as cross entropy. Should the IIT objective be fully met, then $\mathcal{M}_{\mathcal{H}}, \mathcal{N}^{\theta}$, and Π satisfy the causal abstraction relationship as in Equation 3.1.

In our setting, we do not have a value for $\texttt{INTINV}(\mathcal{M}_{\mathcal{H}}, \mathbf{b}, \mathbf{s}, V)$ as in Equation 4.1, since we do not have an explicit definition of Y in our high-level causal graph. Hence, we modify the IIT objective to a semi-supervised, contrastive learning setting as in the original CLIP training paradigm [48].

Our partial causal model does not provide a deterministic function that computes the effect of P on Y. Nevertheless, it does encode direction: in the case that X is a text input

with $P = \text{capt.}$, then the counterfactual $X_{P \leftarrow \text{desc.}}$ should have a higher similarity score for the same image. Conversely, a counterfactual caption text should receive a lower similarity score for an image than its original text description for the same image. This contrast allows us to set up IIT within a contrastive learning framework, as shown in Figure 4.3. We contrast between CLIP run on the base text input, and CLIP run on the same input but with an interchange intervention with a counterfactual P (i.e. description $\rightarrow$ caption, or vice versa). Our training objective function can be defined as

$$\sum_{\mathbf{b},\mathbf{s},\mathbf{I} \in \mathbf{V}_{in}} \texttt{CE}(\mathbf{I}^T \left[\mathcal{N}^\theta(\mathbf{b}), \texttt{INTINV}(\mathcal{N}^\theta, \mathbf{b}, \mathbf{s}, V)\right], l) \tag{4.2}$$

where $\mathcal{N}^\theta$ is the CLIP text encoding model, initialized with pretrained CLIP weights, $\mathbf{I}$ is a static image encoding, computed by the CLIP image encoder (kept frozen during training), and l is a label with $l = 0$ if $\mathbf{b}$ is a description and $l = 1$ if $\mathbf{b}$ is a caption.

With our modified IIT objective function, we induce a high-level causal structure where the underlying purpose of a text (i.e., caption versus description) has a causal effect on the similarity of the text to the image. In our experiments, we evaluate whether our IIT-CLIP model is sensitive to the communicative purpose of a text, and whether it is otherwise comparable to the original CLIP model.

4.3 Experiment Details

We compare our IIT-CLIP model to the pretrained CLIP model, which achieves strong performance in a variety of image captioning and description tasks. We also compare IIT training to finetuning CLIP with the objective of preferring text descriptions over captions when describing the same image. We evaluate the models on the following tasks, increasing in ambition: (1) ability to distinguish between captions and descriptions in the held-out Concadia test set, (2) correlation of similarity scores to human evaluations of image descriptions, and (3) performance on a zero-shot image classification dataset from a new domain.

The first evaluation serves as a sanity check, ensuring that our trained models are able to distinguish the high-level concept of a text's underlying purpose. The second evaluation follows the hypothesis of [33] that distinguishing descriptions and captions is critical in creating human-centered image descriptions. The third evaluation ensures that the power of

CLIP to match images and texts across various domains is not significantly reduced by the training method. A model that sufficiently passes all three evaluations is likely to serve as a useful model for image description generation.

4.3.1 Description-Caption Distinction

We use the test set of Concadia to evaluate whether CLIP distinguishes between captions and descriptions. Across all (image, caption, description) triples in the held-out test set, we compute the rate at which the CLIP score for the (image, description) pair is higher than the CLIP score for the (image, caption) pair. Since we are fitting the CLIP model for the task of evaluating image alt-description generation models, a higher rate of descriptions over captions is preferred.

4.3.2 Correlation to Human Evaluation

We follow the work of [32] to evaluate the correlation between CLIP scores and human evaluations of text descriptions for images. We obtain CLIP scores for 70 image-text pairs extracted from Wikipedia, and compute their correlation with human evaluations, both sighted and BLV individuals, for these same image-text pairs [32]. We hypothesize that a model which better distinguishes between descriptions and captions will also align better with human judgement.

4.3.3 Domain Transfer

Lastly, we seek to ensure that a finetuned CLIP model does not compromise CLIP's strong performance on cross-domain transfer. This is important for the sake of description generation across varying distributions of images. We evaluate domain transfer by computing the model's accuracy on CIFAR-100, a zero-shot image classification task with 100 image classes [34]. For a given image, we predict its class by choosing the text description of highest CLIP similarity score to the image, in the form of "An image of <class>". Our goal is to find a training paradigm which balances the trade-off between channeling CLIP to distinguish between descriptions and captions while maintaining CLIP's zero-shot capabilities in a previously unseen image domain.

The use of causal abstraction towards accessible image descriptions can be viewed as a form of addressing bias with respect to real-world principles: although multi-modal models

Model	Desc. > Capt.	Accuracy
CLIP	49.4%	61.4%
+ IIT	84.3%	46.3%
+ finetuned	89.0%	40.9%

Table 4.1: Results table for description vs. caption preference, and transfer to CIFAR-100. **Desc. > Capt.** reports the rate at which, for a description-caption pair describing the same image, the model assigns a higher similarity score to the description-image pair than the caption-image pair. **Accuracy** reports accuracy on the CIFAR-100 dataset, where a random classifier has an expected accuracy of 1 %.

such as CLIP do not distinguish the purpose of a text, we as humans hypothesize that this distinction is critical for providing accessible image descriptions. By inducing a causal abstraction relationship between CLIP and our high-level causal model, we (1) produce an interpretable model whose computation with respect to text purpose is human-understandable, and (2) align the model's high-level computation path with human understanding of the world. Hence, inducing causal abstractions forms a connection between interpreting language models and addressing their biases. In the next section, we investigate the case when a language model aligns with an undesired causal structure, and how to reduce the causal effect of its intermediate variable.

4.4 Results

Table 4.1 reports the results of (1) a multi-modal model's ability to distinguish between captions and descriptions within the Concadia held-out test set [33]; and (2) a multi-modal model's ability to perform zero-shot image classification. A higher **Desc. > Capt.** percentage means that the model prefers description texts over caption texts within a counterfactual pair of texts that describe the same image – hence aligning with human understanding of a text's communicative purpose in an image description task. A higher **Accuracy** means that the model achieves high description accuracy in a separate image domain – hence preserving the original capabilities of CLIP.

As previously shown by Kreiss et al. [33], CLIP does not distinguish between descriptions and captions, preferring descriptions to their counterfactual captions 49.4% of the time. Hence, we wish modify CLIP to instill this description-caption distinction, while preserving its original image-to-text capabilities. We acknowledge that finetuning CLIP is just as, if not more, effective than interchange intervention training in terms of instilling the distinction

		Overall	Imaginability	Relevance	Irrelevance
BLV	CLIP	0.08	0.10	0.09	0.09
	+finetuned	0.22	0.22	0.24	0.17
	+IIT	0.11	0.26*	0.29*	0.26*
Sighted, no image	CLIP	-0.01	0.06	0.00	-0.17
	+finetuned	0.14	0.11	0.13	-0.04
	+IIT	0.25*	0.24*	0.21	0.05
Sighted, with image	CLIP	0.14		0.11	-0.08
	+finetuned	0.09		0.09	0.06
	+IIT	0.30*		0.26*	0.18

Table 4.2: Correlation between model similarity scores and human preference, across imaginability, relevance, irrelevance, and overall value of text description [32]. Scores reported with an asterisk (*) are statistically significant with $p < 0.05$.

between descriptions and captions within CLIP (89.0% by finetuning vs. 84.3% by IIT). Nevertheless, IIT is more robust to task transfer than finetuning (40.9% by finetuning vs. 46.3% by IIT), and hence preserving the original image-to-text performance of CLIP across different image domains.

This is potentially because unlike the finetuning objective, which trains a model to prefer descriptions over captions with respect to the Concadia dataset, the IIT objective explicitly localizes the high-level concept of communicative purpose. Hence, finetuning is more prone to overfit on the training image distribution, whereas IIT aligns CLIP with a more robust interpretable causal structure. We posit that the balance struck by IIT between debiasing CLIP and maintaining its original task performance is most useful for alt-description generation tasks.

The correlation between model outputs and human preference, reported in Table 4.2, further supports our hypothesis that IIT aligns CLIP with human understanding of alt-descriptions and their purpose. CLIP trained with IIT is the only model to achieve statistically significant correlation with human preference. Interestingly, only in the overall score for BLV participants does finetuning achieve a higher correlation than IIT (0.22 by finetuning vs. 0.11 by IIT). Yet on all other metrics rated by BLV participants (i.e., imaginability, relevance, and irrelevance), IIT not only outperforms finetuning, but also achieves statistically significant correlation. We posit that since IIT distinguishes the communicative purpose of the text, it is better at capturing the relevant details of an image, without adding in details that are only present in the context. Hence, although the overall valuation of BLV individuals more strongly correlates with finetuned CLIP than IIT-CLIP,

the valuation of IIT-CLIP is more grounded in the visual details of the image. The strong correlation between IIT-CLIP and the relevance score by sighted participants who were given access to the image, in stark contrast to the finetuned relevance score, supports this hypothesis (0.26 by IIT versus 0.09 by finetuning).

5. Reducing Causal Effect: Debiasing Language Models

As language models improve in understanding and replicating the nuanced structure of natural language, they also risk amplifying and perpetuating the bias of its speakers. In this section, adapting the technique of causal abstraction, we introduce a light-weight model editing method for achieving counterfactual fairness with respect to a sensitive attribute. We demonstrate our debiasing method in a supervised setting (i.e., sentiment analysis) and an unsupervised setting (i.e., text completion) with respect to gender. In both settings, our method reduces bias while maintaining the model's original capabilities, outperforming a recently developed causal intervention method as well as a commonly-used debiasing technique.

5.1 Gender Bias Evaluation for Language Models

Although gender bias in language models is ubiquitous and well-documented, there is no centralized resource towards evaluating models for gender bias and mitigating its effects [56] (see Section 2.1 for more detail). In this thesis, we investigate and attempt to mitigate gender bias in a supervised sentiment analysis setting, and an unsupervised text generation setting. We limit our definition and evaluation of gender bias to the Equity Evaluation Corpus (EEC) [29] and the Professions datasets [59, 38, 4].

The EEC dataset consists of gender-swapped sentence pairs, such as "The man is happy" and "The woman is happy." By evaluating the difference of a language model's output on these sentence pairs, we can estimate the counterfactual behavior of a model should the gender of the sentence subject be swapped. We utilize the EEC dataset to evaluate and mitigate bias across gendered subject noun phrases in language models.

The Professions dataset consists of incomplete clauses as prompts for a generative language model, such as "The doctor said that" and "The nurse said that." We note that such sentences do not in themselves constitute a counterfactual pair – the doctor and nurse professions differ by more than just their stereotypically associated gender. Nevertheless,

with enough imprecise pairs such as "The doctor said that" and "The dancer said that", we hypothesize that the profession information will be averaged out, and the key distinction will be the stereotypically associated gender. This perspective bears similarity to the meta-sampling approach in [61], which interprets sentiment analysis language models trained on restaurant reviews.

We acknowledge that our debiasing experiments are limited by the definition and evaluation of gender bias supported by our choice of datasets. For example, the EEC dataset simplifies gender to a binary variable (e.g., "he" vs. "she", "man" vs. "woman"). Additionally, gender may be implicitly encoded within natural language without an explicit gendered noun phrase. Likewise, by evaluating the probability of a pronoun token in describing the profession noun in the sentence, the Professions dataset does not support a spectrum of gender identities. Our experiments with the Professions dataset are restricted to the male pronoun "he", female pronoun "she", and gender neutral pronoun "they". We do not expect our debiased method to generalize to the full continuum of gender identities nor the full extent to which gender underlies social interactions. However, we believe that given increasingly inclusive and detailed dataset, our method has the potential to decreasingly biased language models. We provide a thorough analysis of our method's limitations in Section 6.

5.2 Distributed Alignment Search

Whereas in Section 4 we induce a causal effect between the high-level concept of communicative purpose and a multi-modal model's behavior, here we seek to reduce the causal effect between the high-level concept of gender and a language model's behavior. Hence, rather than predefine an alignment and induce a desired causal abstraction, we search for an alignment that surfaces an undesired causal abstraction. We find this alignment using distributed alignment search (DAS), a recently developed method for uncovering causal abstractions [20]. Then, we introduce a method for intervening on a low-level language model so as to reduce the undesired causal effect of a high-level causal variable, and approach counterfactual fairness.

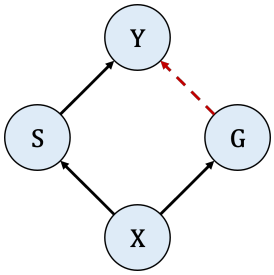


Figure 5.1: High-level causal model for the purposes of debiasing a sentiment analysis language model with respect to the high-level concept of gender.

5.2.1 Our Structural Causal Model

Suppose that we are given a language model biased with respect to gender, as defined by either the EEC or Professions datasets. How can we evaluate and surface the bias of our model? In our method, we begin by seeking to interpret our language model. We do so via hypothesizing an undesired causal structure – one where the gender of the main noun phrase has a causal effect on the model's output – and determining whether it is an accurate causal abstraction of our biased language model.

Let us focus on the supervised sentiment analysis setting. Our high-level casual model $\mathcal{M}_{\mathcal{H}}$ consists of four causal variables: X, which represents an input sentence, and whose value can be any natural language sentence; S, which represents the level of joy in the input sentence, and whose value can be any real number; G, which represents the gender of the subject of the input sentence, and whose value can be either M or F; and Y, which represents the output of our language model, and whose output can be any real number. This high-level causal model is visualized in Figure 5.1. Finally, let $\mathcal{N}$ be the frozen parameters of the language model we would like to debias. We explicitly define the value of Y in our high-level causal model as $F_Y = \mathcal{N}(X)$. That is, the output of our high-level causal model reflects the original bias of our low-level language model.

We note that $\mathcal{M}_{\mathcal{H}}$ reflects the gender bias we would like to remove. Nevertheless, we seek to find an alignment Π that reflects a causal abstraction relationship between $\mathcal{M}_{\mathcal{H}}$ and $\mathcal{N}$. Why should we want to guarantee a causal abstraction between a language model and our undesired high-level causal model? Our insight is that a causal abstraction between $\mathcal{M}_{\mathcal{H}}$ and $\mathcal{N}$, should it exist, would serve two purposes: (1) we can interpret the bias of our language model with respect to the high-level causal model $\mathcal{M}_{\mathcal{H}}$, and (2) we can construct interventions with respect to this causal abstraction that will reduce the undesired causal

effect between gender and the language model's output. We provide a brief outline of the DAS algorithm, which we use to find an alignment that asserts this causal abstraction, before discussing our intervention method for debiasing our language model.

5.2.2 Distributed Alignment Search (DAS)

The DAS method is a lightweight method for finding a causal alignment without modifying the original parameters of a neural network [20]. Given a predetermined activation layer within a neural network, DAS learns an invertible orthogonal projection (i.e., rotation matrix) from the neural activations of that layer to a latent "high-level concept" space. Given a counterfactual pair of inputs, DAS computes the orthogonal projection of the two hidden activation layers, performs an interchange intervention within the "high-level concept" space, and then inverses the orthogonal projection back to the original activation space. The rotation matrix is optimized to meet the interchange intervention training objective, as in Equation 4.1. By using a projection matrix, DAS can uncover latent causal concepts that may be encoded within a different basis [20]. By enforcing an orthogonal projection, DAS maintains the property of associativity between high-level concepts, which is essential for causal abstraction [18].

We apply DAS to find an alignment between our high-level causal structure $\mathcal{M}_{\mathcal{H}}$ and biased language model $\mathcal{N}$, and we evaluate our alignment using interchange intervention accuracy. Yet unlike in Section 4, a causal abstraction does not imply a debiased model – in fact, it simply reasserts that our existing model is biased. How can we intervene on our language model in order to remove the causal effect of gender on its output?

5.3 Intervention for Causal Abstraction

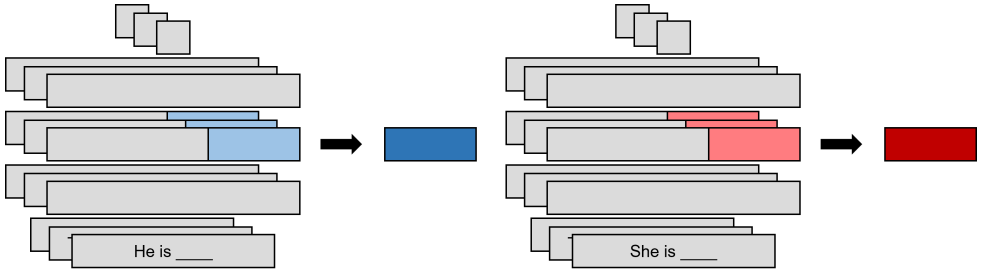
Let us revisit our example biased fact-checking model one last time. In Section 3.4, we saw that given an alignment Π between our high-level causal model $\mathcal{M}_{\mathcal{H}}$ and the biased language model $\mathcal{M}_{\mathcal{L}}$ that satisifes the causal abstraction relationship Equation 3.1, we can intervene on our low-level model $\mathcal{M}_{\mathcal{L}}$ in order to debias it with respect to the degree of confidence of its input sentence. We did this by choosing an arbitrary vector that fit the size of our alignment, and intervening on the computation of $\mathcal{M}_{\mathcal{L}}$ using that vector and our alignment Π. We apply the same idea to mitigate gender bias in language models.

We saw that, when the causal abstraction is consistent across all inputs, it is enough to choose any intervention vector in order to achieve counterfactual fairness. However, in practice, the interchange intervention accuracy is less than one-hundred percent. This means that our intervention vector serves two purposes: (1) reduce the causal effect of the high-level concept of gender as much as possible, and (2) maintain the original performance of the biased language model in all other contexts. In this thesis, we explore methods for fixing an intervention vector, as well as learning an intervention vector to satisfy our constraints.

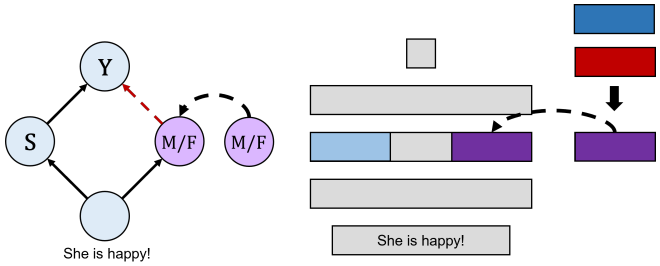
5.3.1 Fixing an Intervention

One idea is that our intervention vector should "null out" the high-level concept of gender within the language model's computation. To implement this, we can construct a vector $\mathbf{v}_{null} = \vec{0}$, which replaces each neural activation that correspond to the high-level causal variable G with zero. Though this intervention is likely to reduce the effect of the causal variable G on the resulting language model, it might compromise the original model's capabilities.

Another idea is that our intervention vector should capture the "average" of our high-level concept of gender, as encoded within the language model's computation [61]. We visualize this approach in Figure 5.2. First, we compute the neural activations of our language model $\mathcal{N}$ when run on sentences with male noun phrases. We use our learned alignment Π to extract the activations that correspond to the high-level variable G in our causal graph $\mathcal{M}_{\mathcal{H}}$. Similarly, we compute the neural activations when run on sentences with female noun phrases, and extract the values that correspond to the high-level variable G. By averaging our male-sentence activations $\mathbf{v}_M$, we have a single averaged "male" vector that abstracts the high-level concept of $G = M$; correspondingly, we can compute an averaged vector $\mathbf{v}_F$ for the high-level concept $G = F$. By intervening on our model with the averaged male intervention, $\mathcal{N}_{\Pi(G) \leftarrow \mathbf{v}_M}$, we can expect to answer the question, "what would the output of the model have been, if the main noun phrase of the sentence was male?" Yet what if we were to intervene on the model with an averaged $\mathbf{v}_{avg} = (\mathbf{v}_M + \mathbf{v}_F)/2$ vector representation? This thesis hypothesizes that such a fixed vector strikes a balance between reducing the effect of gender on a language model's output while maintaining its performance in other contexts.



(a) Constructing an average intervention. By extracting the neural activations of a language model at the site aligned to the high-level concept of gender, we can hope to capture a "prototype" gender intervention. The outputted intervention vector in blue is an averaged $G \leftarrow M$ intervention, while the vector in red is an averaged $G \leftarrow F$ intervention.



(b) By intervening on our low-level language model with an averaged intervention (here, an average of the $G \leftarrow M$ and $G \leftarrow F$ vector), we hope to remove the causal effect between the language model's low-level implementation of G and its final output.

Figure 5.2: Example fixed intervention computed as an average of a $G \leftarrow M$ "prototype" intervention and the corresponding $G \leftarrow F$ intervention. We hope that this process removes the causal connection between G, the high-level concept of gender, and Y, the output of our low-level language model.

5.3.2 Learning an Intervention

Our thesis also considers learning an intervention vector. To learn an intervention vector, we can redefine our causal structure by setting $F_Y = S$, meaning that Y no longer causally depends on G (i.e., removing the dashed red arrow in Figure 5.1). We can compute S as the average sentiment score assigned by the model between a male and female counterfactual pair (e.g., the average sentiment of "He is happy" and "She is happy"). Now, the interchange intervention training objective, Equation 4.1, ensures that our learned intervention vector $\mathbf{v}_\theta$ does not stray away from the original model predictions. Since we learn a fixed intervention vector across all inputs, we still guarantee counterfactual fairness as in Definition 3.

By defining some intervention $\mathbf{v}$ and intervening on a biased language model $\mathcal{N}$, we produce a language model $\mathcal{N}_{\Pi(G)\leftarrow\mathbf{v}}$ that is both interpretable with respect to causal abstraction and fair with respect to counterfactual fairness. Through investigating and mitigating gender bias in language models, we demonstrate the connection between interpretability

and fairness in practice.

5.4 Experiment Details

We evaluate our intervention for causal abstraction method in two settings: a supervised sentiment analysis setting, and an unsupervised text generation setting.

5.4.1 Supervised Setting: Bias in Sentiment Analysis

We investigate gender bias in a pretrained BERT model [12], finetuned on a sentiment analysis benchmark, SemEval 2018 [42]. The finetuned BERT model places within the top ten models, out of more than two hundred submissions, on the SemEval 2018 benchmark. Nevertheless, as measured by the Equity Evaluation Corpus (EEC) [29], BERT demonstrates significant gender bias in its outputted sentiment score.

We construct a high-level causal structure to audit BERT for gender bias, and confirm a causal abstraction using the counterfactual data from the Equity Evaluation Corpus (EEC). Utilizing the alignment between BERT and our posited high-level causal structure, we develop intervention vectors to reduce the causal effect of gender on the predicted sentiment score.

We compare our technique to iterative nullspace projection (INLP) [50], as well as to the original model. We report (1) gender bias, as measured by EEC, and (2) score on the original SemEval task. That is, we seek to measure the trade-off between (1) achieving counterfactual fairness and (2) maintaining strong performance on the supervised sentiment analysis task.

5.4.2 Unsupervised Setting: Bias in Natural Language Generation

We evaluate gender bias in a pretrained GPT-2 model [49]. As demonstrated by the Professions bias evaluation dataset, GPT-2 reflects bias in associations between professions and gender stereotypes [59]. Utilizing DAS and intervention selection, we attempt to mitigate gender bias in GPT-2 with respect to the pronoun associated with a given profession.

We compare our technique to Rank-One Model Editing (ROME) [40], a causal editing method based on interpretability through causal tracing [59]. We report (1) gender bias, as measured by the total effect of a profession's gender stereotype on the pronoun used

Model	Bias (% M > F)	Bias (M - F)	Accuracy
BERT	90.4%	3.82	0.692
+ INLP	58.5%	0.82	0.280
+ Intervention (M)	68.8%	0.76	0.647
+ Intervention (F)	83.9%	1.42	0.644
+ Intervention (Null)	66.4%	0.89	0.658
+ Intervention (Avg.)	75.1%	1.10	0.646

Table 5.1: Results EEC dataset (supervised setting). Bias is measured as the percentage rate of preferring a certain gender to another within a counterfactual pair (**% M > F**), and the average absolute difference in model outputs (**M - F**), as measured by EEC [29]. Accuracy is reported as a correlation score between the model outputs and human labels on the SemEval 2018 held-out test set [42].

by GPT-2 on the Professions dataset; (2) natural language understanding, as measured by the GLUE score [60]; and (3) the rate of correspondence between the pronoun and the profession in the model-generated sentence, hand-labeled across 30 example sentences. That is, we seek to measure the trade-off between (1) achieving counterfactual fairness and (2) maintaining the capabilities of the original GPT-2 model. However, we believe that the total effect of a profession's gender stereotype on the generated pronoun is an insufficient metric for bias. Often, a model might output the anti-stereotypical pronoun, but to index a different entity (e.g., "the doctor$_i$ said that she$_j$ will heal soon"). Hence, we measure (3) the "depth" of the underlying gender bias in GPT-2.

5.5 Results

5.5.1 Supervised Setting: Bias in Sentiment Analysis

Table 5.1 presents results for our experiment in debiasing a pretrained BERT model [12] that is finetuned on the SemEval 2018 sentiment analysis training set [42]. We find that our method significantly reduces bias, without compromising the original model's performance. However, we also note that not all fixed interventions operate identically, and recommend future research on intervention generation.

Our debiasing setting assumes access to (1) a pretrained BERT model, finetuned on a sentiment analysis task, and (2) the EEC dataset, with respect to which we would like to debias BERT. Note that we do not assume access to the original SemEval training set. This is because in most settings, off-the-shelf models do not have readily accessible training data, and re-training language models can be computationally expensive. Nevertheless,

we measure both (1) the bias of our outputted model, with respect to EEC, and (2) the performance of our model, with respect to the held-out SemEval test set. Hence, our setting encourages finding a lightweight model debiasing technique that preserves the original model's capabilities without reiterating on their original training process.

We find that a BERT model, finetuned on a sentiment analysis task, scores 0.692 out of 1.00 on the SemEval benchmark, placing within the top 10 models evaluated on this benchmark. Yet BERT displays statistically significant bias by assigning higher joy sentiment scores to sentences with a male noun phrase over the counterfactual sentences with a female noun phrase 90.4% of the time. We confirm this bias using DAS. We find an alignment between BERT and our hypothesized biased causal model with an interchange intervention accuracy of 87.7%, suggesting that in almost all inputs, the gender of the main noun phrase is computed by BERT and plays a causal role in its output.

Our method successfully debiases BERT without completely compromising its original performance on this task. In particular, our null intervention (an all-zeros vector) is unbiased with respect to gender as measured by EEC, assigning higher scores to male inputs 66.4% of the time and achieving an average difference of 0.89. Furthermore, the null-intervention model scores 0.658 on the SemEval task, placing within the top 20 models (out of more than 200 models) evaluated on this task. Hence, our experiment confirms the potential of the interpret-and-intervene method towards causally debiasing language models.

Interestingly, we find that not all interventions have the same consequences. For example, although an averaged M intervention achieves low bias (68.8% male-over-female preference), the average F intervention results in a model that is still quite biased (83.9% male-over-female preference). This is likely because our alignment does not guarantee as a perfect causal abstraction relationship. Rather, Equation 3.1 holds for 87.7% of counterfactual input pairs from EEC. We hypothesize that our averaged female intervention falls into the distribution of inputs where Equation 3.1 no longer holds. To address this in future research, we encourage work on searching for alignments with high interchange intervention accuracy, as well as work on selecting the most appropriate intervention based on a given alignment.

5.5.2 Unsupervised Setting: Bias in Natural Language Generation

Table 5.2 presents results for our experiment in debiasing a natural language generation GPT-2 model. We compare our intervention method to ROME [40], a method for locating

Model	Surface Bias	Perplexity	Underlying Bias
GPT-2	244.4	15.69	0.47
+ ROME	34.6	14.82	0.60
+ Intervention (M)	8.7	14.80	0.93
+ Intervention (F)	16.5	14.80	0.93
+ Intervention (Null)	8.3	14.80	0.90
+ Intervention (Avg.)	12.8	14.80	0.90

Table 5.2: Results for the Professions dataset (unsupervised setting). Surface bias is measured as the total effect between the associated stereotype of a profession and the respective probabilities of generating "he" versus "she" pronouns [59]. Perplexity is measured on the held-out Wikitext dataset [41]. Underlying bias is the rate at which the pronoun generated by the underlying model coreferences the profession in the original prompt.

and editing factual associations within language models.

We employ DAS to search for an alignment between GPT-2 and a partial high-level causal model that is biased with respect to gender, as in Figure 5.1. We find an alignment that achieves an interchange intervention accuracy of 0.76, suggesting that GPT-2 can be largely abstracted by this biased causal model. Hence, we expect that intervening on the aligned high-level concept of gender, as in Figure 5.2b, will remove the causal effect of gender on the model-generated text.

Indeed, we find that our intervention approach significantly reduces the total effect of gender on GPT-2 output, as measured by [59]. Our method achieves a lower total effect than the ROME method [40] (8.7 versus 34.6, respectively). One reason for this is that ROME induces factual associations in the form of "profession" $\rightarrow$ "gender pronoun" (e.g., "doctor" $\rightarrow$ "she"), whereas our method abstracts the high-level concept of gender across all professions. This means that our method better generalizes to new professions and their associated stereotypes.

Our results confirm that our method reduces gender bias in GPT-2 without compromising its original performance in text generation. Our debiasing method achieves a perplexity score on the held-out Wikitext 2 dataset [41] that is comparable to the perplexity score of GPT-2 (14.80 versus 15.69, respectively), meaning that our debiased model generates fluent text. Although we require more experiments to confirm this, we hypothesize that our debiased model can replace the original GPT-2 model in the original natural language generation settings of GPT-2.

Our results demonstrate that our method (1) reduces the original gender stereotype bias of GPT-2 on the Professions dataset, and also (2) achieves comparable performance to

GPT-2 in natural language generation. Nevertheless, we emphasize that our model does *not* constitute a debiased version of GPT-2. As elaborated upon in Section 6, the Professions dataset measures a single and somewhat limited dimension of gender bias. We do not expect our method to generalizes to dimensions of gender bias outside of equalizing the rate of pronouns that are generated following a particular profession within a sentence.

We evaluate one dimension of bias, which we term "underlying bias" in Table 5.2, on a random sample of 30 templates from the Professions dataset. On a given generated text output, we manually decide whether the pronoun in the generated sentence coreferences the profession in the template, or an external noun entity. For example, consider the sentence "The doctor said that she will recover quickly." Although the sentence uses the antistereotypical pronoun "she", this pronoun refers to the patient, and not the doctor. Our surface bias evaluation does not capture this nuance, and so we report it manually as an underlying bias. We find that our method, despite reducing surface bias, does not remove the underlying bias in GPT-2 (underlying bias rate of 0.90). We encourage future work on bias evaluation methods for deeper underlying bias in language models. We believe that, as bias evaluation methods increase in complexity and nuance, so too will our method result in more deeply debiased models.

6. Limitations

Our work proposes a connection between interpretability and algorithmic fairness, and demonstrates the potential of this connection in debiasing language models. However, we emphasize that our experiments are, for the moment, purely for demonstration purposes. Here we elaborate on the current limitations of our experiments, and point out points of caution for research that seeks to apply our debiasing technique.

Our experiments are limited by access to bias evaluation datasets and mechanisms. For example, our gender debiasing experiments apply a binary definition of gender, which by itself perpetuates harmful gender bias in the form of underrepresentation of non-binary gender definitions [56, 64]. This is because our bias evaluation methods, EEC and the Professions dataset, are restricted to a binary definition of gender. Additionally, our unsupervised debiasing setting focuses on the pronoun generated by a language model, but does not consider the overall sentiment, implication, or quality of the outputted sentence with respect to that pronoun.

For example, we evaluate our models on the prompt "The doctor was promoted because...", expecting debiased models to output "she" and "he" at an equal rate. Our GPT-2 model modified by ROME completes the sentence as follows: "The doctor was promoted because she had a baby,' said a neighbor who did not want to be named." Although the model uses the antistereotypical pronoun "she" to refer to the doctor, the outputted sentence encodes deep gender bias and perpetuates a denigrating concept of female professionals [10]. Our simplified bias evaluation system does not capture this deep and nuanced gender bias. As language models increase in their depth of language understanding, we must develop richer bias evaluations that surface deep underlying bias. We believe that, given richer evaluation metrics, our method can remove deeper levels of bias in existing language models.

We also acknowledge a key ontological problem in causal reasoning with respect to protected attributes, which applies directly to our debiasing method. In particular, in designing a high-level causal model, we must ask "What do its variables *mean*?" Recent literature on causal reasoning points out that the meanings of social constructs such as gender, race, or religion might not ever be fully captured by a single causal variable [28]. Social

constructs have a cyclical causal effect between their perception, (i.e., the effect of constructs on their surroundings), and their formation story (i.e., the effect of the surroundings on the definition of the construct) [25]. Causal models, which are directed acyclic graphs by definition [45], cannot capture this cyclic relationship. One way in which this thesis addresses this limitation is by constraining the meaning of a social construct, such as gender, to its effect on the biased model itself, with respect to a bias evaluation dataset. In this sense, the gender of a noun phrase within a sentence has a causal effect on the language model, while the language model does not contribute to the definition of the high-level concept of gender. As language models become integrated into online resources and communities, this approach may no longer effective – one can foresee a future in which social constructs such as gender, race, or religion are partly defined by the same language models that we are seeking to debias with respect to these constructs.

7. Conclusion and Future Work

In this thesis, we assert a theoretical connection between causal interpretability and counterfactual fairness. Utilizing causal abstractions and a new intervention generation method, we debias state-of-the-art language models with respect to real-world principles (i.e., that a text's communicative purpose plays a role within image description tasks) and ethical principles (i.e., that the gender of a sentence's subject should not affect the overall sentiment of the sentence). We hope that future research will build atop of our method by post-processing existing language models to generate interpretable, unbiased language models with similar capabilities.

The field of NLP is uniquely situated in that its technology is far better at generating realistic-looking solutions than evaluating said solutions. We encourage the field of NLP to place the evaluation of language models – for bias, toxicity, fairness, understanding, general intelligence, sentience, or even consciousness – as a top priority. We hope that this thesis provides a direction towards designing such evaluation processes for bias, fairness, and interpretability, that future research will undertake.

Bibliography

[1] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021.

[2] Jack M Balkin and Reva B Siegel. The american civil rights tradition: Anticlassification or antisubordination. *Issues in Legal Scholarship*, 2(1), 2003.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[6] Michael Brownstein. Implicit Bias. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition, 2019.

[7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[8] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[9] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL*

Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 276–286, Florence, Italy, August 2019. Association for Computational Linguistics.

[10] Kate Crawford. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*, 2017.

[11] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103, 2015.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

[14] Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108, 2021.

[15] Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

[16] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[17] Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586, 2021.

[18] Atticus Geiger, Chris Potts, and Thomas Icard. Causal abstraction for faithful model interpretation. *arXiv preprint arXiv:2301.04709*, 2023.

[19] Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable

neural networks. In *International Conference on Machine Learning*, pages 7324–7338. PMLR, 2022.

[20] Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D Goodman. Finding alignments between interpretable causal variables and distributed neural representations. *arXiv preprint arXiv:2303.02536*, 2023.

[21] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[22] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

[23] Bernease Herman. The promise and peril of human evaluation for model interpretability. *arXiv preprint arXiv:1711.07414*, 2017.

[24] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[25] Daniel Hirschman and Isaac Ariail Reed. Formation stories and causality in sociology. *Sociological Theory*, 32(4):259–282, 2014.

[26] Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926, 2018.

[27] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online, July 2020. Association for Computational Linguistics.

[28] Atoosa Kasirzadeh and Andrew Smart. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 228–236, 2021.

[29] Svetlana Kiritchenko and Saif M Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*, 2018.

[30] Ezra Klein. The surprising thing a.i. engineers will tell you if you let them. *The New York Times*, 2023.

[31] Gerd Kortemeyer. Could an artificial-intelligence agent pass an introductory physics course? *arXiv preprint arXiv:2301.12127*, 2023.

[32] Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. *arXiv preprint arXiv:2205.10646*, 2022.

[33] Elisa Kreiss, Noah D Goodman, and Christopher Potts. Concadia: Tackling image accessibility with context. *arXiv preprint arXiv:2104.08376*, 2021.

[34] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[35] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[36] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[37] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631, 2013.

[38] Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing. *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pages 189–202, 2020.

[39] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172, 2013.

[40] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.

[41] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[42] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17, 2018.

[43] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015.

[44] Ji Ho Park, Jamin Shin, and Pascale Fung. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*, 2018.

[45] Judea Pearl. *Causality*. Cambridge university press, 2009.

[46] Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[47] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[49] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[50] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*, 2020.

[51] Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online, April 2021. Association for Computational Linguistics.

[52] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, June 2016. Association for Computational Linguistics.

[53] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[54] Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. On second thought, let's not think step by step! bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061*, 2022.

[55] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[56] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

[57] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[58] Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational*

Linguistics, pages 4593–4601, Florence, Italy, July 2019. Association for Computational Linguistics.

[59] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, 33:12388–12401, 2020.

[60] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[61] Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. Causal proxy models for concept-based model explanations. *arXiv preprint arXiv:2209.14279*, 2022.

[62] Gal Yona and Guy Rothblum. Probably approximately metric-fair learning. In *International Conference on Machine Learning*, pages 5680–5688. PMLR, 2018.

[63] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

[64] Lal Zimman. Transgender language, transgender moment: Toward a trans linguistics. 2018.