This WACV 2023 Workshop paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Causal Structure Learning of Bias for Fair Affect Recognition

Jiaee Cheong
Computer Science & Technology
University of Cambridge, UK
jc2208@cam.ac.uk

Sinan Kalkan
Computer Engineering
Middle East Technical University, Turkey
skalkan@metu.edu.tr

Hatice Gunes
Computer Science & Technology
University of Cambridge, UK
hatice.gunes@cl.cam.ac.uk

Abstract

The problem of bias in facial affect recognition tools can lead to severe consequences and issues. It has been posited that causality is able to address the gaps induced by the associational nature of traditional machine learning, and one such gap is that of fairness. However, given the nascency of the field, there is still no clear mapping between tools in causality and applications in fair machine learning for the specific task of affect recognition. To address this gap, we provide the first causal structure formalisation of the different biases that can arise in affect recognition. We conducted a proof of concept on utilising causal structure learning for the post-hoc understanding and analysing bias.

1. Introduction

The problem of bias in machine learning tools is becoming a greater source of concern. This is also true for the task of facial affect recognition as such tools are increasingly deployed in a wide-range of tasks ranging from from medicine [13] to driver drowsiness detection [28]. The problem of bias is compounded by the fact that the machine learning algorithms are often black-box in nature. This not only hampers their wide-spread adoption but also makes it harder to understand the reason or source of the biases and to tackle them. Many works have discussed the challenge of non-human interpretable nature of current machine learning models and have sought ways to address them [32].

Causal reasoning tools are specifically designed to tease out the underlying causal mechanisms and has been proposed as a potential instrument to address such gaps [18, 22, 27]. First, it is important to highlight the distinction between conventional causal inference and causal inference as applied to big data. The former can be understood as a suite of methods comprising of statistical mechanisms coupled with the usage of directed acyclic graphs (DAGs). Research is predominantly centered upon causal inference (effect estimation) and structure learning (causal pattern discovery). The latter is more frequently associated with methods that combine other machine learning algorithms with causal reasoning tools to address the limitations of existing machine learning (ML) methods. Therein lies the first challenge: there is still a need to map from causal inference methods to causal machine learning. In addition, though research in algorithmic fairness is rapidly expanding, most of the existing works are tailored towards and benchmarked against social or tabular datasets [23]. No existing literature formalises the different types of bias prevalent in facial affect recognition setups in terms of causal graphs. A principled framework for accounting for bias in affect recognition is still missing.

Our contributions can be summarised as follows. First, we provide the first formalisation of the prevalent types of bias in affect recognition using causal graphs. We posit that it is only possible to causally-debias outputs in a principled manner only if we have accounted for the right causal pathways that induce such biases in the first place. Second, we provide a proof of concept of a post-hoc pattern search method that can be used to understand bias. Third, we provide an analysis of our findings and highlight the existing opportunities and challenges. Section 2 reviews the relevant literature in the field. Section 3 provides some technical preliminaries necessary to understand structural causal models and causal structure learning. Section 4 formalises the different types of bias and their respective representation in terms of causal graphs. Section 5 provides the research methodology. We analyse the results in Section 6 and provide further discussions in Section 7.

2. Literature Review

2.1. Fairness in Facial Affect Recognition

Facial affect recognition involves methods that attempt to analyse and predict facial affect [35]. There are different ways to do so. One prevalent method is to describe expressions as discrete categories. Paul Ekman and his colleagues proposed that there are six basic emotion categories of facial expressions (i.e. happiness, surprise, fear, disgust, anger and sadness), that are claimed to be recognised universally [11]. Another way to analyse facial affect is by us-

ing the Facial Action Coding System (FACS), a taxonomy of human facial expressions in the form of Action Units (AUs) [11]. Other facial affect description include representing affective states as bipolar entities which exists on a continuum [34].

At present, the investigation of bias in facial affect recognition is still very much an understudied problem [5,30,45]. There is only a smattering of studies which attempted to analyse the bias and propose fairer solutions for facial affect recognition [3, 8, 16, 28, 45]. In addition, most of the literature do not take causal relations into account [8, 16, 28, 45]. It is only in recent times that more studies have attempted to leverage causality to address the problem of bias in facial affect recognition [3, 6].

2.2. Causality for Facial Affect Recognition

Causal inference has proven to be highly effective at tackling several computer vision based tasks such as image captioning [47], semantic segmentation [49] and few-shot learning [48]. A natural extension of the above would be to attempt to leverage causal inference for another computer vision task: facial expression recognition. Indeed, existing attempts at doing so have been highly successful [2, 29]. However, existing works have only leveraged on sequential data input [29] or investigated the use of interventions, back-door adjustments, and confounders [2, 3]. None of the existing works have explored the usage of structural causal models to formalise bias. Oh et al. [29] introduced a modular causality extractor which is independent of the feature extractor. Though no specific causal mechanism was utilised, the authors imposed a "causal relation" by virtue of learning the relationship between past facial images to current affect states. On the other hand, Chen et al. [2] addressed the AU recognition subject variation problem by removing the confounding effect caused by the *confounder* 'Subject'. Chen et al. [3] addressed the dataset bias problem by proposing a network to induce backdoor adjustment in order to deconfound the dataset-related context features such as background scenes from the target emotion feature.

Our work differs from existing work in several crucial ways. First, no existing work have investigated the usage of structural causal models (SCMs) and causal structure discovery methods for the problem of bias and fairness in facial affect recognition. We provide the first formalisation of bias in affect recognition using SCMs. Second, existing methods have attempted to identify bias based on metric-based evaluations. We attempt to identify bias using causal pattern search algorithms. Third, most of the bias mitigation strategies do not distinguish between the different source of *path-specific bias*. For instance, with reference to Figure 1, there is no method that distinguishes between the bias that originates from the sensitive attribute A or the bias that originates from the context Z. Our method attempts to distinguish between the different source of bias. To date, there is still no formal definitions of the structural causal pathways of the different types and sources of bias. First, we attempt to address the research gap by defining the different types of bias using SCMs. We subsequently illustrated the usage of causal discovery methods to provide post-hoc identification and analysis of the potential source of bias.

3. Causality: Technical Preliminaries

3.1. Structural Causal Models

In this paper, variables are denoted by capital letters. A is used for the sensitive attribute (e.g. race, gender). $\hat{Y}$ is used as the predicted outcome. A Structural Causal Model (SCM) [31] M is defined as a triplet (U, V, F), with U, V and F sets defined in the following manner:

1. U is a set of latent background or exogeneous variables which affect the model but yet are not represented within the model.

2. $V = \{V_1, ..., V_n\}$ is the set of observable or endogeneous variables within the model.

3. F is the set of functions $\{f_1, ..., f_n\}$, one for each $V_i \in V$, such that $V_i = f_i(pa_i, U_{pa_i})$, $pa_i \subseteq V \backslash \{V_i\}$, $U_{pa_i} \subseteq U$.

The notation "pa_i" refers to the "parents" of V_i and is motivated by the assumption that the model factorizes as a directed graph, here assumed to be a directed acyclic graph (DAG). The model is "completely specified" when both instantiations $U = u$ and F are given.

3.2. Causal Structure Learning

Causal structure learning aims to infer a causal model from data. Causal models not only describe the observational joint distribution of variables but also formalize predictions under interventions and counterfactuals [31, 38]. Directed acyclic graphs (DAGs) are commonly used to represent causal structure: Nodes represent variables and directed edges point from cause to effect representing the causal relationships. This graphical representation rests on assumptions which have been critically questioned, for example by Dawid [9]. Inferring causal structure from observational data is non-trivial. Often, we can only identify the DAG up to its Markov Equivalence Class (MEC) and finding high-scoring DAGs is NP-hard [7]. The implication of the above is that we can only discover causal patterns up to MEC. The MEC limitation compounded with the independence assumptions therefore determines the edges or arrows that are present within a "discovered" causal graph. Table 2 provides a summary of this.

We can also understand DAGs as graphical models that represent a set of hypotheses about the causal process that

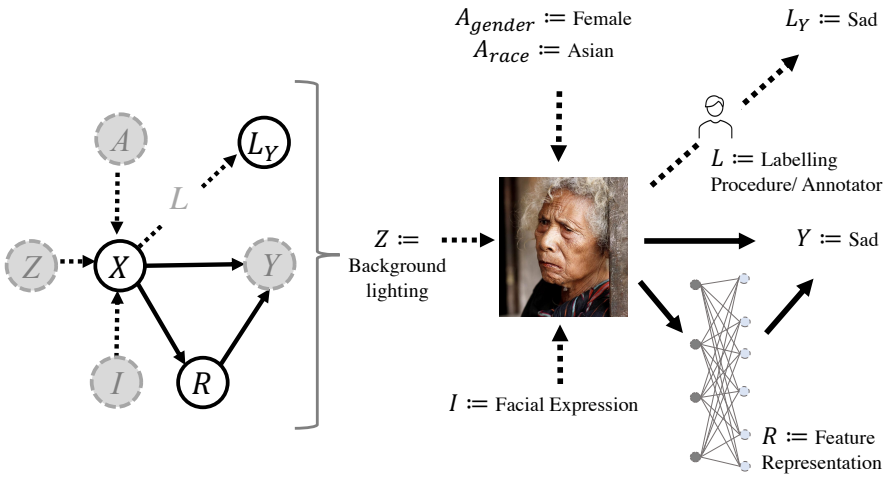

Figure 1. Black nodes and arrows correspond to the observed variables and observable functions. Grey nodes correspond to the unobservable variables and functions. $X :=$ image, $R :=$ image representation learnt by the machine learning model, $Y :=$ ground-truth labels, $L_Y :=$ class labels labelled by the annotator, $A :=$ ground truth sensitive attributes (e.g., race, gender, age), $Z :=$ contextual factors or background noise and $I :=$ image features which should be directly relevant towards the true class label Y.

engendered the set of variables of interest. Edges embody the hypothesised conditional dependencies. Conversely, unconnected nodes represent variables that are conditionally independent of one another. The arrow $X \rightarrow Y$ illustrates a potentially direct causal effect of X on Y. This means that Y is directly influenced by X. Hence, altering X via external interventions would also affect Y. It is important to note that an arrow $X \rightarrow Y$ only represents a causal effect which is unmediated by any other variables within the graph. The arrow is omitted if it is certain that X does not have a direct causal effect on Y. There is a variety of Causal Structure Discovery algorithms or methods covered in [44]. In our experiments, we deployed a causal structure learning algorithm called the Fast Causal Inference (FCI) [38]. We will cover the implementation in further detail in Section 5.2.

4. Formulating Bias in Affect Recognition Causally

This section outlines our first contribution. As noted by the numerous surveys on bias and fairness [5, 25], there are many different types of bias. It is important to distinguish between the different types of bias as different forms of bias would necessitate different mitigation strategies. Note that not all of the different biases reviewed in existing surveys [25] map to bias in facial affect recognition [5]. For instance, there is no equivalence for the problem of recidivism for facial affect recognition. In order to address the problem of bias for facial affect recognition from a causal perspective, we first attempt to delineate between the different biases present and its implication on the resulting causal graph. The novelty here is that no existing work have formalised the different types of bias for affect recognition using causal graphs.

Let the random variables X and Y represent the images and their respective labels, Given an input image $X = x$, the goal of the image classification task is to predict its label, $Y = y$. Assuming a statistical probabilistic interpretation, both X and Y are presumed to follow the following conditional probability distribution:

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X). \quad (1)$$

A typical machine learning approach constructs a learner to learn $P(Y|X)$ given X and Y. During classification, it will then pick a class that satisfies $\arg\max_y P(Y = y|X = x)$.

Note that the above formulation is purely statistical. To formulate the above from a causal perspective, we will make use of the SCM as explained in Section 3. This SCM is represented in both the left causal graph in Figure 1. The black nodes are the observable variables whereas the grey nodes are the unobservable variables. The solid arrows correspond to observable causal relationships (i.e., with well-defined parameterised functions) whilst the dashed arrows correspond to non-observable causal relationships given the data. As the annotation or labelling process (denoted by L) is never fully observable or transparent, we have opted to capture this using dashed arrows. Though it is possible to estimate or recover the parameters of this function using observed data, we can never be entirely certain of its ground truth generative process. With reference to the example on the right in Figure 1, the nodes represent the following quantities: X represents the image, R represents the image representation learnt by the machine learning model, Y represents the ground-truth labels , L_Y represents the class la-

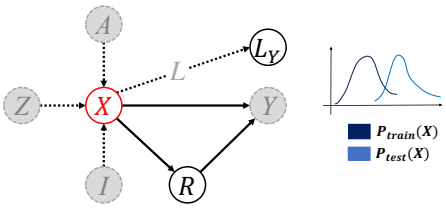

Figure 2. Dataset bias can largely be understood as the bias arising from node X as highlighted in red. However, as the other variables A, Z and I are unobserved, it is difficult to evaluate the precise source of bias.

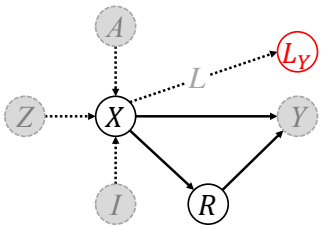

Figure 3. Labelling bias L_Y can largely be understood as the bias due to the labeling process L. This is a source of bias that is distinct from other computer vision and tabular data settings.

bels labelled by the annotator, A represents the ground truth sensitive attributes (e.g., race, gender, age), Z represents the contextual factors or background noise and I represents the image features which should be directly relevant towards the class label Y. Given the above, X, R, L_Y therefore correspond to V, the set of observed variables, whereas A, Z, I and Y correspond to U, the set of unobserved variable encoding external sources of variation. Note that the set of U listed is non-exhaustive. In addition, between every variable, the arrows represent the set of mechanisms F which defined the functional relation between the variables. For instance, $X \leftarrow f_X(A, Z, I, U)$ and $L_Y \leftarrow f_{L_Y}(X, U)$ where U represents unobserved variables encoding other sources of variations not captured in Figure 1.

4.1. Dataset Bias

In general, a dataset is deemed unbiased if the joint distribution $P_{model}(X, Y)$ (or $P_{train}(X, Y)$) matches that of reality $P_{reality}(X, Y)$ (or $P_{test}(X, Y)$). The gap between the two distributions may largely be deemed to be an domain shift problem. Most debiasing techniques for image or facial affect recognition focus on the bias from the image data X [20, 42]. This can be understood as the bias that arises from the images themselves $P(X)$. This form of bias is illustrated in Figure 2 as represented by the red node X.

Intuitively, this means that the distribution parameters (i.e., the means, standard deviation, skews etc.) that characterise the population differ across demographic subgroups. We define subgroups as subsets of the population defined across certain sensitive attributes (e.g., race, age and gender). With reference to Figure 2, if we only have information about the statistical distribution parameters, dataset bias in the form of distribution mismatch does not tell us much as we would not know whether the source of bias is A, Z or I. Identifying the correct source of bias is crucial towards picking the right mitigation methods. For instance, without visual inspection, we are unable to ascertain whether a mismatch between P_{train} and P_{test} is due to I or Z. If the source of mismatch is due to I, (i.e., there is an imbalanced representation between the image features relevant towards the prediction task) the way to address this would be to make sure the dataset distribution of P_{train} and P_{test} matches. However, other mitigation methods would be needed if the true source of bias is the context Z. In this instance, domain generalisation methods would be a more suitable mitigation method instead [8].

4.2. Labelling Bias

Debiasing $P(X)$ is not the ultimate panacea. With reference to Equation 1, we see that $P(X, Y)$ will still be biased if the annotated labels $P(Y|X)$ are biased. Statistically, in order for Equation 1 to hold, this would therefore require the annotated labels $Y|X$ to be unbiased. Labelling or annotation bias can be understood as the bias stemming from the labelling process. Labelling bias is a form of bias more prevalent or observed in affect recognition. This is due to the subjective nature of facial affect expression and recognition. In a conventional computer vision setup, there is typically no or minimal discrepancy between L_Y and Y as they are typically the same. For instance, assuming an object detection task where the object in question is a car. Compared to an affect recognition task, the discrepancy between Y and L_Y will be lesser as in most cases, we can usually collectively and objectively agree on whether the object is in fact a car or not. However, this is more challenging for affect recognition.

First, there is the discrepancy between self-reported, third-party and machine or algorithm-labelled annotations [4]. Existing research on affect recognition has indicated that depending on the way affect is labelled (intended, self-reported and observed), the outcome or accuracy of an algorithm can differ widely [46]. In addition, there is another layer of bias introduced by the individual labellers. Taking third-party annotation as an example, across gender, it has been noted that third-party observers are likelier to perceive females faces as happier than males [39]. Across race, there is the problem of the Other Race Effect (ORE). ORE corresponds to the phenomena where individuals are often better at recognising people of their own-race than they are of the other-race [26]. The hypothesised explanation is that the ability to do so is a result of having more experience of discriminating among people from a homogeneous group of similar face (i.e., faces of an individual's own race) [41]. We see similar evidence for affect recognition as well [12, 14].

Table 1. RAF-DB Emotion label distribution breakdown percentage across Gender, Race and Age. The values in bold represents the emotion class percentage that differs the most from the overall percentage of the subgroup within the sample.

	Gender		Race			Age					
Emotion	Male	Female	Cauc	AA	Asian	$0-3$	$4-19$	$20-39$	$40-69$	$70+$	Percent.
Surprise	46.5%	53.5%	87.5%	5.4%	7.1%	12.1%	13.1%	60.6%	12.1%	2.0%	10.3%
Fear	54.4%	45.6%	77.2%	6.3%	16.5%	3.8%	8.9%	63.3%	20.3%	3.8%	2.7%
Disgust	43.7%	56.3%	79.1%	3.8%	17.1%	1.9%	8.2%	67.1%	17.7%	**5.1%**	5.5%
Happy	37.6%	**62.4%**	74.9%	8.6%	16.5%	3.8%	18.9%	50.9%	23.1%	3.2%	39.7%
Sad	38.1%	61.9%	75.4%	7.8%	16.8%	**13.2%**	**25.1%**	42.5%	15.8%	3.4%	13.4%
Angry	**72.6%**	27.4%	87.8%	6.1%	6.1%	1.2%	9.8%	70.1%	17.1%	1.8%	5.7%
Neutral	47.9%	52.1%	75.1%	6.0%	18.9%	3.1%	12.7%	**70.4%**	**10.4%**	3.4%	22.6%
Percent.	43.7%	56.3%	77.4%	7.1%	15.5%	5.5%	16.4%	57.5%	17.4%	3.2%	

However, existing machine-learning based bias mitigation methods have often treated this form of systematic bias as a source of random noise which are unbiased on average [50] even though affect-based literature has cast doubt on such assumptions [12, 14, 39]. It is only in recent years that computer vision methodologies have acknowledged and recognised this [4]. Recent works in computer vision have proven labelling bias to be non-random. It is a systematic form of bias which can be mitigated if properly accounted for [4]. However, the shortcoming is that metrics-based solutions which typically attempts to quantify or illustrate the presence of bias using score-based methods (e.g. equality in accuracy) are still susceptible of leading towards trivial prediction score adjustment rather than a fundamental bias reduction [36]. Intuitively, this form of bias is evidenced by the graph in Figure 3. There is bias if L_Y differs substantially from Y. There is no or minimal bias if L_Y aligns with Y. We can identify and mitigate this if we have lab-collected data where the ground truth Y is provided by the actual subjects. However, for images collected in-the-wild, we typically only have access to L_Y but not Y.

4.3. Contextual Bias

Here, we use the term contextual bias as the overall background environment (e.g. image backdrop or contextual scene) that an image is placed in. This is a pernicious cause of disparity in algorithm performance and is widely investigated as an out-of-distribution generalisation problem in computer vision [24]. For the specific computer vision problem of facial affect recognition, the biggest dichotomy is between that of a "lab-controlled" and an "in-the-wild" dataset. Other sources of contextual bias include pose, lighting, outdoor versus indoor, camera equipment etc. [40]. This is a non-trivial form of bias. For instance, research have shown that the image quality explained some, but not all, of the variation in algorithm performance across race [43]. The findings indicated that when ICAO-compliant (i.e., good quality) images where used, accuracy improved across board.

With reference to Figure 4, the source of bias is captured

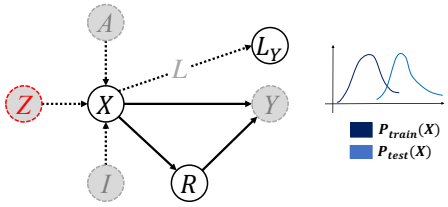

Figure 4. Contextual bias represented by node Z is subsumed within X and is thus very difficult to tease out. Statistically, this translates into a mismatch in distribution between P_{train} and P_{test} as captured by the graph on the right. It is difficult to distinguish the true cause unless we have access to highly controlled lab-settings.

by node Z. However, Z is generally unobservable as it is subsumed within X. At first glance, this may seem as a rather trivial observation but it does facilitate strong claims about the generative process and dictates the efficacy of existing methods. With the causal diagram captured in Figure 4, we are claiming that contextual bias is a source of bias that is part of $P(X)$, which makes it harder to tease out unless we are relying on highly-controlled lab settings or simulated/ generated images. Note that just by analysing the distribution, we will be unable to determine where the mismatch in distribution is due to change in context or a genuine dataset bias. While this may not be true in all experimental setups, we believe that this is true for many settings especially when working with natural images. Getting the causal effect for natural or in-the-wild images is extremely challenging. This is because there are innumerable *unobervable* confounding factors within real-world data.

5. Research Methodology

With these preliminaries in place, we now empirically and analytically explore the utility of causal structure-learning for the problem of affect recognition. To illustrate the themes detailed in Section 4, we study a simple setup for facial affect classification. The study can be divided into two main stages. First, we train a basic black-box prediction model, a ResNet-18 on the raw images to predict the emotion class of each image. Subsequently, we run the Fast

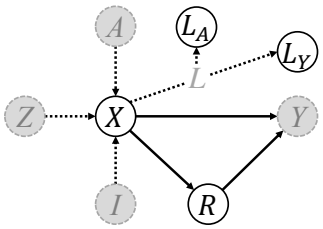

Figure 5. Causal graph of the existing experimental setup. Grey nodes represent variables that are not observed whilst black nodes represent variables that are observed.

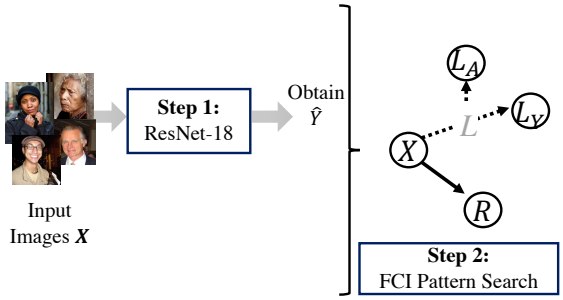

Figure 6. Experimental setup and causal graph with exogeneous variables removed. In this setup, there is no direct association between any of the variables.

Causal Inference (FCI) [38] algorithm to learn the causal structure between the variables, predicted labels and actual labelled emotions. Figure 5 represents the ground truth causal relationship between the variables. In order to facilitate assessment, we removed all the unobserved variables to arrive at an ecosystem which reflects the environment that we are running our causal structure learning algorithm over. This is represented by Figure 6.

5.1. Dataset, Pre-processing and Model Training

We conducted our experiments on the RAF-DB dataset [21]. It is a real-world dataset curated from the Internet. The images are manually annotated with expression and sensitive attribute labels. RAF-DB contains labels in terms of facial expressions of emotions (surprise, fear, disgust, happy, sad, anger and neutral) and sensitive attribute labels along gender, race and age. We excluded images labelled as "unsure" for gender. We utilised a subset of the dataset consisting of 14,388 images. 11,512 samples were used for training and 2,876 samples were used for testing. This training and testing split has been pre-defined according to the instructions in the original dataset [21]. It is available for non-commercial research purposes and researchers are able to gain access to it by contacting the authors [21]. A breakdown of the dataset distribution is illustrated in Table 1.

All images are cropped to ensure faces appear in relatively similar positions. The images are then normalized to a size of 128×128 pixels and fed into the networks as input. During the training stage, we apply the following commonly used augmentation methods: Randomly cropping the images to a slightly smaller size (i.e., 96×96); rotating them with a small angle (i.e., range from -15^o to 15^o); and horizontally mirroring them in a randomized manner. ResNet-18 [15] is used in our experiments. ResNet-18 and the experimental setup was chosen to align with the existing research in this area [6, 45]. In addition, ResNet-18 also provides good performance-time trade-off. We used the PyTorch implementation of ResNet. We trained it from scratch with the Adam optimizer [19], with a mini-batch size of 64, and an initial learning rate of 0.001. The learning rate decays linearly by a factor of 0.1 every 40 epochs. The maximum training epochs is 100, but early stopping is applied if the accuracy does not increase after 30 epochs.

5.2. Causal Structure Learning: FCI

We have chosen to use the Fast Causal Inference (FCI) algorithm [38] for our experiment. This is because the typical pattern search algorithm, such as the PC algorithm [38], assumes causal sufficiency (i.e, that there are no unmeasured common causes) which is not pragmatic given our settings. The FCI algorithm works similarly to the PC algorithm but relaxes the assumption of causal sufficiency. The result is therefore known as a partial ancestral graph (PAG). The FCI predominantly consists of two phases:

1. **An adjacency phase:** The adjacency phase of the algorithm starts with a complete undirected graph.

2. **Orientation phase:** FCI then enters an orientation phase that uses the stored conditioning sets (that previously led to the removal of adjacencies) to orient as many of the edges as possible.

The specific steps or pseudo code can be found in [38]. In order for the FCI algorithm to account for latent common causes of variables, the PAG adds a circle symbol that can be placed at either end of an edge in the same way an arrowhead can be. As a result, the different edges should be interpreted differently. Table 2 provides a summary of this.

Prior or background knowledge is a pre-imposed set of conditions where certain variables cannot or must cause others. The constraints can also be different variables that are in different time orders, and thus cannot be the cause of one another. Forbidden graphs specify the causal relationships that are not allowed in the eventual causal model. In our setup, the forbidden relationships include the following. First, we prohibit any relation from emotion labels to other variables such as age, race and gender. Second, we also prevent causal links between standalone sensitive attributes by the common sense that they are independent of one another.

Table 2. Different types of edges and their causal interpretations.

Edge Types	Potentially Present Causal Relationships	Absent Relationships
$A \rightarrow B$	A is a cause of B. However, the causation may either be direct or indirect (i.e. there exists other variables along the pathway). In addition, there is potentially an unmeasured confounder between A an B.	B is not a cause of A.
$A \leftrightarrow B$	There exists an unmeasured confounder U between A and B. In other words, there may be variables along the causal pathways from U to A or from U to B.	A is not a cause of B. B is not a cause of A.
A o$\rightarrow B$	Either A is a cause of B (i.e.: $A \rightarrow B$) or there is an unmeasured confounder between A and B (i.e.: $A \leftrightarrow B$) or both.	B is not a cause of A.
A o–o B	Exactly one of the following holds: 1. A is a cause of B 2. B is a cause of A 3. There is an unmeasured confounder of A and B. 4. Both 1. and 3. 5. Both 2. and 3.	-
Bold or thickened edges	There is no latent confounder. Otherwise, latent confounders might be present.	
Green edges	If an edge is green, the relationship is certainly direct. Otherwise, it is only *possibly* direct.	

6. Experiments and Results

We conducted two sets of experiments. The FCI algorithm will attempt to discover the potential causal relations between the variables available (L_Y, L_A and R). With reference to Figure 6, we see that with the given variables, the experiments will only be able to help us gain insights over the labelling process L which relates to the labelling bias discussed in Section 4.2. The first set corresponds to an environment or setup where R is not included, i.e., the setup mainly involves the sensitive attributes and ground truth emotion label. The second set corresponds to the environment or setup where R is included, i.e., the setup includes the sensitive attribute, ground truth emotion label and emotion class predicted by the black-box learner. For each analysis, we compare the causal structure learnt when background or prior knowledge is supplanted vs not.

6.1. Causal Structure Learning with no Predictions

First, we analyse the causal structure learnt when we do not include the black-box predictions within the ecosystem. With reference to Figure 7, we see that the causal structure learnt is not very informative. With reference to Table 2, we know that, with an edge with two circles X o-o Y, we cannot even guarantee an adjacency, i.e., there is no set that d-separates X and Y. Hence, the graph on the left in Figure 7 indicates that we have not managed to learn any causal relation without supplanting prior knowledge. The graph on the right is slightly more informative. Given A o$\rightarrow B$, the only conclusion is B is not an ancestor of A. Hence, we know that *Emotion Label* is not an ancestor of *Age*, *Race* and *Gender*. Though the causal structure learnt after supplanting prior knowledge (where we have imposed certain prohibitions as explained in Section 5.2) may seem slightly more informative, note that the oriented edges are mainly a result of us supplying prior information.

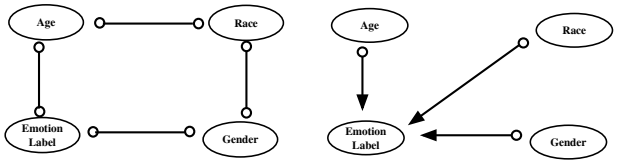

Figure 7. **Causal structure learnt without providing black-box predictions:** (a) The graph on the left is the causal structure learnt without supplanting prior knowledge as discussed under Section 5.2. (b) The graph on the right is the causal structure learnt after supplanting prior knowledge.

6.2. Causal Structure Learning with Predictions

Next, we analyse the causal structure learnt when we do include the black-box predictions within the ecosystem. With reference to Figure 8, we see that this produced more informative results than before. Analysing the graph on the left, we see that there is an unmeasured confounder between the variables *Age* and *Gender* and between the variables *Gender* and *Emotion Label*. Interestingly, we see a green

arrow from *Age* to *Emotion Label* and from *Emotion Label* to *Emotion Predictions*. If an edge is green, the relationship is most likely direct. The interpretation here is that *Age* is certainly a direct cause of *Emotion Label* and *Emotion Label* is certainly a direct cause of *Emotion Predictions*.

Analysing the graph on the right after supplanting prior knowledge, we see that *Race* is independent other variables. This is noteworthy as the omission of an arrow is a stronger claim than the inclusion of an arrow. This is because the presence of an arrow depicts merely the "causal null hypothesis" that X might have an effect on Y. Though the two o$\rightarrow$ arrows between *Age* and *Emotion Label* and *Gender* and *Emotion Label* may seem more informative, as before, this is merely a result of us supplanting prior information before performing the FCI analysis. A reassuring aspect is that the green arrow from *Emotion Label* to *Emotion Predictions* still stands.

In addition, the results suggest that there might be some form of labelling bias across the sensitive attributes *Gender* and *Age*. Since we have the o$\rightarrow$ arrow between *Age* and *Emotion Label* as well as *Gender* and *Emotion Label*, this means that across both *Age* and *Gender*, they might either be a cause of *Emotion Label* or that there is an unmeasured confounder between them. Indeed, looking at Table 1, across *Gender*, we see a wider disparity between the emotion class breakdown and that of the subgroup breakdown. For instance, there is approximately 43.7% males and 56.3% females. An emotion class that is generally faithful to this breakdown is the category "Surprise" where we have 46.5% males and 53.5% females. However, the same is not true across "Happy" and "Angry". As we can see, there is disproportionately more females categorised as "Happy" and disproportionately more males categorised as "Angry". This trend is not observed across *Race* but we do see a similar trend across *Age*. There can be two reasons for this. First, it may be a class imbalance problem where there is more male samples for the emotion "Angry". As such, the images X in Figure 6 would be the confounding factor. The second plausibility is that the labelling process, denoted by L in Figure 6 is a confounding factor for both the labelled sensitive attributes L_A and labelled emotion class L_Y. Despite the inherent limitations of causal structure learning as discussed in Section 3, we are still able to gain some post-hoc understanding of the bias present.

7. Conclusion

According to the proof-of-concept conducted and preliminary assessment of the results, it is reassuring that the annotated *Emotion Label* L_Y from the dataset seems to be the only variable that has a direct cause on the *Emotion Predictions* outputted by the black-box (ResNet) classifier. Our results hints that there may potentially be labelling bias as discussed in Section 4.2.

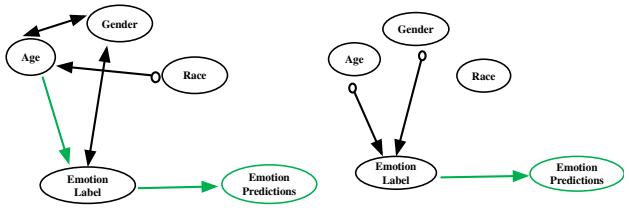

Figure 8. **Causal structure learnt after providing black-box predictions:** (a) The graph on the left is the causal structure learnt without supplanting prior knowledge. (b) The graph on the right is the causal structure learnt after supplanting prior knowledge.

An opportunity is to deploy causal structure learning for algorithm auditing in affect modelling or any other biometrics-related technology. Causal learning will equip the community with the tools needed to account for the various theoretical and conceptual aspects of affect recognition. For instance, studies indicate that different genders express (and perhaps experience) emotion differently [1, 10]. Similarly, emotion recognition accuracy differs across cultures [17, 33, 37]. However, a limitation is that causal structure learning on its own may not be sufficient. This is because once a pattern or a PAG has been produced from data, any single DAG that we select from the equivalence class will be equivalent in its causal structure. The causal structure itself has little use if not combined with statistical parameters and prior knowledge about the data generation process. Another limitation is that there is insufficient evidence to scientifically explain each source of bias. Part of this is due to the lack of balanced datasets for this problem setting. There is currently no affect-based dataset (e.g. CK+, AffectNet) that is balanced across classes. We hope this work will encourage other researchers to explore the usage of SCMs and causal pattern search to build fairer and more robust models.

Acknowledgements

Open access statement: For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.
Data access statement: This study involved secondary analyses of pre-existing datasets. All datasets are described in the text and cited accordingly. Licensing restrictions prevent sharing of the datasets. The authors thank Shan Li, Prof Weihong Deng and Jun-Ping Du from the Beijing University of Posts and Telecommunications (China) for providing access to RAF-DB.
Acknowledgement: J. Cheong is supported by the Alan Turing Institute doctoral studentship and the Cambridge Commonwealth Trust. H. Gunes' work is supported by the EPSRC under grant ref. EP/R030782/1.

References

[1] Tara M Chaplin. Gender and emotion expression: A developmental contextual perspective. *Emotion Review*, 7(1):14–21, 2015.

[2] Yingjie Chen, Diqi Chen, Tao Wang, Yizhou Wang, and Yun Liang. Causal intervention for subject-deconfounded facial action unit recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[3] Yuedong Chen, Xu Yang, Tat-Jen Cham, and Jianfei Cai. Towards unbiased visual emotion recognition via causal intervention. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 60–69, 2022.

[4] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14980–14991, 2021.

[5] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. The hitchhiker's guide to bias and fairness in facial affective signal processing: Overview and techniques. *IEEE Signal Processing Magazine*, 38(6):39–49, 2021.

[6] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Counterfactual fairness for facial expression recognition. *2022 ECCV Workshop on Challenge on People Analysis (WCPA)*, 2022.

[7] Max Chickering, David Heckerman, and Chris Meek. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

[8] Nikhil Churamani, Ozgur Kara, and Hatice Gunes. Domain-incremental continual learning for mitigating bias in facial expression and action unit recognition. *IEEE Transactions on Affective Computing*, 2022.

[9] A Philip Dawid. Causal inference without counterfactuals. *Journal of the American statistical Association*, 95(450):407–424, 2000.

[10] Yaling Deng, Lei Chang, Meng Yang, Meng Huo, and Renlai Zhou. Gender differences in emotional response: Inconsistency between experience and expressivity. *PloS one*, 11(6):e0158666, 2016.

[11] Rosenberg Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[12] Hillary Anger Elfenbein and Nalini Ambady. *Is there an in-group advantage in emotion recognition?* American Psychological Association, 2002.

[13] Y. Gurovich, Y. Hanani, O. Bar, G. Nadav, N. Fleischer, D. Gelbman, L. Basel-Salmon, P. Krawitz, S. Kamphausen, M. Zenker, L. Bird, and K Gripp. Identifying Facial Phenotypes Of Genetic Disorders Using Deep Learning. *Nature Medicine*, 2020.

[14] Jennifer N Gutsell and Michael Inzlicht. Empathy constrained: Prejudice predicts reduced mental simulation of actions during observation of outgroups. *Journal of experimental social psychology*, 46(5):841–845, 2010.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Ayanna Howard, Cha Zhang, and Eric Horvitz. Addressing bias in machine learning algorithms: A pilot study on emotion recognition for intelligent systems. In *2017 IEEE Workshop on Advanced Robotics and its Social Impacts (ARSO)*, pages 1–7. IEEE, 2017.

[17] Shunhang Huang, Junjie Qiu, Peiduo Liu, Qingqing Li, and Xiting Huang. The effects of same-and other-race facial expressions of pain on temporal perception. *Frontiers in Psychology*, 9:2366, 2018.

[18] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *NIPS*, pages 656–666, 2017.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

[20] Shan Li and Weihong Deng. A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing*, 2020.

[21] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593. IEEE, 2017.

[22] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

[23] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*, 2022.

[24] Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. Causal transportability for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7521–7531, 2022.

[25] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2019.

[26] Christian A Meissner and John C Brigham. Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1):3, 2001.

[27] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[28] M. Ngxande, J. Tapamo, and M. Burke. Bias remediation in driver drowsiness detection systems using generative adversarial networks. *IEEE Access*, 8:55592–55601, 2020.

[29] Geesung Oh, Euiseok Jeong, and Sejoon Lim. Causal affect prediction model using a past facial image sequence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3550–3556, 2021.

[30] Jaspar Pahl, Ines Rieger, Anna Möller, Thomas Wittenberg, and Ute Schmid. Female, white, 27? bias evaluation on data and algorithms for affect recognition in faces. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 973–987, New York, NY, USA, 2022. Association for Computing Machinery.

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

[32] Mohit Prabhushankar and Ghassan AlRegib. Extracting causal visual features for limited label classification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3697–3701. IEEE, 2021.

[33] B Nicole Reyes, Shira C Segal, and Margaret C Moulson. An investigation of the effect of race-based social categorization on adults' recognition of emotion. *PloS one*, 13(2):e0192418, 2018.

[34] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178, 1980.

[35] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(6):1113–1133, 2015.

[36] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Information-theoretic bias reduction via causal view of spurious correlation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[37] Jose Angel Soto and Robert W Levenson. *Emotion recognition across cultures: the influence of ethnicity on empathic accuracy and physiological linkage.*, volume 9. American Psychological Association, 2009.

[38] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[39] John Eric Steephen, Samyak Raj Mehta, and Raju Surampudi Bapi. Do we expect women to look happier than they are? a test of gender-dependent perceptual correction. *Perception*, 47(2):232–235, 2018.

[40] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, 3(1):16–30, 2021.

[41] Diana Su Yun Tham, J Gavin Bremner, and Dennis Hay. The other-race effect in children from a multiracial population: A cross-cultural comparison. *Journal of experimental child psychology*, 155:128–137, 2017.

[42] Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In *Domain adaptation in computer vision applications*, pages 37–55. Springer, 2017.

[43] Kushal Vangara, Michael C King, Vitor Albiero, Kevin Bowyer, et al. Characterizing the variability in face recognition accuracy relative to race. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[44] Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *ACM Computing Surveys (CSUR)*, 2021.

[45] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision*, pages 506–523. Springer, 2020.

[46] Hao-Chun Yang and Chi-Chun Lee. Annotation matters: A comprehensive study on recognizing intended, self-reported, and observed emotion labels using physiology. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–7. IEEE, 2019.

[47] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[48] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *Advances in neural information processing systems*, 33:2734–2746, 2020.

[49] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33:655–666, 2020.

[50] Honglei Zhuang and Joel Young. Leveraging in-batch annotation bias for crowdsourced active learning. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 243–252, 2015.