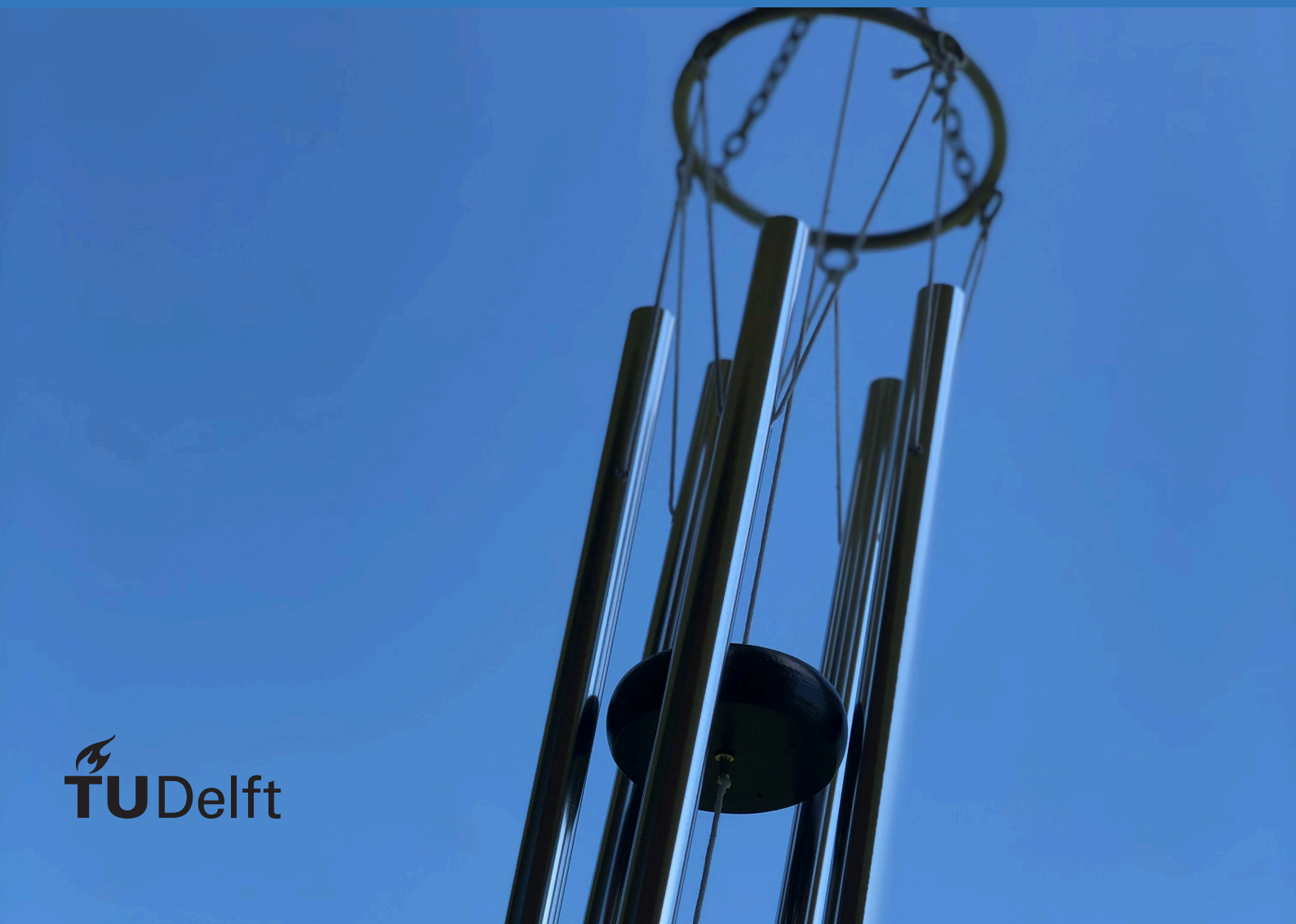
CHIME: Causal Human-In-the-Loop Model Explanations

Shreyan Biswas



TU Delft

CHIME: Causal Human-In-the-Loop Model Explanations

by

Shreyan Biswas

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday August 31, 2020 at 10:30 AM CEST.

This thesis is confidential and cannot be made public until

TU Delft

Abstract

Explaining the behaviour of Artificial Intelligence models has become a necessity. Their opaqueness and fragility are not tolerable in high-stakes domains especially. Although considerable progress is being made in the field of Explainable Artificial Intelligence, scholars have demonstrated limits and flaws of existing approaches: explanations requiring further interpretation, non-standardised explanatory format, and overall fragility. In light of this fragmentation, we turn to the field of philosophy of science to understand what constitutes a good explanation, that is, a generalisation that covers both the actual outcome and, possibly multiple, counterfactual outcomes. Inspired by this, we propose CHIME: a human-in-the-loop, post-hoc approach focused on creating such explanations by establishing the causal features in the input. We first elicit people's cognitive abilities to understand what parts of the input the model might be attending to. Then, through Causal Discovery we uncover the underlying causal graph relating the different concepts. Finally, with such a causal structure, we compute the causal effects different concepts have on a model's outcome. We evaluate the Fidelity, Coherence, and Accuracy of the explanations obtained with CHIME with respect to two state-of-the-art Computer Vision models trained on real-world image data sets. We found evidence that the explanations reflect the causal concepts tied to a model's prediction, both from the perspective of causal strength and accuracy.

Acknowledgements

Over the course of last year working on this thesis was nothing short of a roller-coaster ride. But as they say, it is not about the destination, nor is it about the journey but the people you share your journey with. I have been extremely lucky and blessed in sharing my journey with so many wonderful people, all of whom I want to thank from the bottom of my heart.

The journey began 2 years back in a thought-provoking classroom lecture by Prof. G.J.P.M. Houben who re-ignited my passion for Computer Science.

My daily supervisors, Prof. Jie and Stefan, were a constant feature every Wednesday afternoon, challenging and pushing me towards coming up with new ideas and not letting me tunnel vision. Needless to say, without those scintillating discussions, their constant motivation, understanding and flexibility this project will not be where it is today. I am humbled at the opportunity to work under such brilliant minds and grateful for the time and effort they put into this project.

Lorenzo, you came into this project later but we immediately connected and due to your amazing ability to adapt to what was a fairly new domain we as a team benefited a lot and achieved new highs.

Agathe, although you were not directly part of this project your work and supervision in courses were a big motivator and I am grateful for it.

Caroline, Rebecca and Ojas you 3 were my pillars. During my darkest hours your never-ending love, kindness and motivation helped me beyond words can express. This project is dedicated to you guys.

Alex, Anitej you guys are the best group mates I have ever had. I am glad I decided to pair up with you in Q3 which then turned into a beautiful friendship. The Dream Team.

Victor, you were my first friend in the Netherlands, our endless discussions about life helped me survive during all those days in lockdown, and you always made me feel at home.

Jie and Haiyin you guys came towards the very end but when you did you gave me many reasons to smile and helped me conclude this amazing journey.

Lastly, I would like to thank my parents for providing me this opportunity to come to the Netherlands and allowing me to pursue my dreams and my sister who has been a constant source of light in my life.

Shreyan Biswas
Delft,

Contents

1

Introduction

Context Curiosity is inherent to being human (Kidd & Benjamin, 2015). Thus when the usage of deep neural networks became the vogue of the decade everyone became curious about its internal workings. But curiosity is not the only reason, the potency of these automated marvels made them ubiquitous to a multitude of societal domains where ethical concerns and lack of trust for the users of these systems called for an increased focus on fairness and interpretability (Miller, 2019). As a byproduct, there were a plethora of research methodologies and a generation of researchers seeking to make these black-box models transparent which resulted in an almost unbridled amount of research output in the domain of explainable artificial intelligence (XAI in short).

Problem Focus As part of this thesis, we focus on explaining the outcome of a computer vision model to end users. The rise of deep neural networks can be primarily contributed to the pioneering domain of computer vision Greene, 2020. Over the years it enjoyed heightened performance across various benchmarks. This, however, came at the cost of transparency (Freitas, 2014). Past research has established that these computer vision models while ubiquitous in various social domains (Lee et al., 2019) show biased behaviours Mehrabi et al., 2019. Recently scientific literature written by Wu and Zhang, 2016 titled ”Automated Inference on Criminality using Face Images” was subjected to massive criticism. Despite their best effort at training their model without representation bias, in a later case study (Bergstrom & West, 2016) it was found that the dataset still had observer bias as images of criminals (from mugshots) and non-criminals (from promotional websites) were sourced differently. Examples like these call for further scrutiny of computer vision models and a paradigm that can explain behaviour to validate the generated outcomes and uncover confounding biases.

Motivation While many attempts were made at explaining deep computer vision models, a fundamental flaw surfaced - The answer to the question ”What makes an explanation an explanation?” - failure to have a proper grounding on the definition of explanation led practitioners of the domain to rely on their intuitions of what constitutes a ‘good’ explanation (Miller, 2019). Woodward, 2003 and Buijsman, 2022 provided insights from the philosophy of science literature exhuming explanatory properties that contribute towards a good explanation. According to them, a good explanation must provide answers to *contrastive why questions* that drill down to presenting a generalisation that covers 1) the factual outcome of the model, and 2) a counterfactual outcome. Presenting factual outcomes, for a computer vision model can be attributed to presenting the salient pixels. Salient pixels are those pixels which contribute most to the outcome of the model. However, lack of semantic clarity in these salient pixels leads to ambiguity when generating an explanation - is the model looking at the colour of the object represented by the pixel cluster or the object itself or perhaps its shape? Answering these questions is fundamental to providing a good explanation, however, it lacks an important property - the consideration for the counterfactual “what-if” scenario. Counterfactual scenarios are those which occur contrary to the factual outcomes. To solidify this let us consider a very common scenario when it comes to discussing algorithmic fairness - In a loan application, the user may get a factual explanation as to why his load was rejected - ”The loan was rejected because your income is less than X”. The end user

in response may query “what-if” my income had been greater than X. Note that the user already has a certain amount of income which constitutes the factual scenario at that time, but wants to query about a counterfactual scenario where his income was higher than his current income. Empirical evidence suggests that the inclusion of these kinds of counterfactual scenarios in an explanation improves user understanding of model behaviour (van der Waa et al., 2018) as it provides more clarity and depth to an explanation.

Contribution Through this work, we aim to establish a properly grounded explainable AI framework that outputs factual explanations of a computer vision model by utilising crowd computing - thereby following the methodology of the people for the people. Furthermore, we introduce the causal framework into the computer vision explainability which apart from establishing causal relationships between semantic concepts and model outcomes, allows the framework the capability to answer what-if questions. As an additional dimension, we also present mediation analysis that showcases how much an explanatory attribute is mediated by another attribute which further clarifies model behaviour.

2

Background Information

This chapter is divided into two main parts. In the first part, concepts that are pertinent to this thesis are introduced. In the second part, scientific research in relevant domains and their corresponding sub-domains are presented and discussed

2.1. Background

In this section, we briefly introduce relevant concepts that are an integral part of this work. The section also mentions terminologies that recur throughout this report.

2.1.1. Computer Vision - Image Classification by Deep Neural Networks

Computer vision is a field within artificial intelligence that focuses on developing techniques that extract and interpret information from visual inputs such as images, and videos to draw meaningful conclusions in various tasks. One of its main goals is to develop autonomous systems that replicate or surpass the behaviour of the human visual system (Huang, 1996). The most common types of algorithms that solve the problems of this domain are called deep neural networks or DNNs. These DNNs are very complex algorithms that contain as many as hundreds of billions of learnable parameters (Brown et al., 2020). Because it is very difficult to gauge their learned behaviour by decoding the entire intricate machinery they are often referred to as black-box models. One subdomain within the field of computer vision is image classification. Here the computer vision algorithms are presented with digital images from which they extract visual cues that help them generate a higher level understanding of the image data (Bezdek et al., 1999) and classify them into specific categories. The most common types of algorithms that do these are convolution neural networks or CNNs (O'Shea & Nash, 2015). Primarily they differ from other deep neural networks in their introduction of convolution layers, these layers contain techniques that help the algorithm learn various visual patterns. These learned patterns often referred to as features are then utilised to help the algorithm categorise the image instance into higher-level representations - called class labels. A class label can be anything with a semantic meaning - from animal names to cancer cell types.

2.1.2. Explainable AI

We briefly mentioned the notation of explainability in AI in Chapter 1, while introducing our work. Here we formally introduce the concept of Explainable AI. Explainable AI undertakes the task of explaining the behaviours of AI systems. Note that the term explainability is often interchanged with interpretability. Defining explainability is difficult and remains a topic of research. However, if we look at its goals and target audience, the notion of XAI becomes clearer.

Firstly, the goal of explainable AI is to provide transparency into a black box model's decision-making the black box into a white box. The ways to achieve this goal can be different and based on these differences XAI frameworks are generally divided into two main categories - 1) **Intrinsic**: Here the objective is achieved by designing a simple model that allows for translucency in its decision-making process. A very popular example of this is decision trees, where the outcome can be predicted based upon the conditions of each node within the tree; 2) **Post-hoc**: These types of methods provide

explainability by scrutinising the outputs generated by the black box model after it has finished training. Note that these categories are not mutually exhaustive as one can utilise intrinsic models for post-hoc explanations. Apart from this, based on applicability the XAI methods can be divided into two more categories - 1) Model-specific and 2) Model-agnostic XAI.

There exists large corpora of diverse sets of machine learning models. Some are very distinctive from others. Due to this, the applicability across XAI frameworks varies. Some methods are more generic whereas others like to focus on a specific category of models. “Model-specific” XAI methods cater to the latter. These methods rely heavily on the internal workings of a model and are more likely to be intrinsic in nature. On the contrary “model-agnostic” methods are more generic and by definition, they do not rely upon a model’s architecture or its trained weights. Instead, they base their work upon post-hoc explainability, *i.e*. by analysing input-output interaction. There exists one more categorisation - based on the scope of the XAI frameworks. The scope here signifies the capacity of the framework to explain the model’s behaviour fully or partially.

Local XAI frameworks focus on generating explanations for specific instances of input. **Global** methods on the contrary aim to explain the behaviour of the entire model. As state-of-the-art deep learning has billions of parameters, aggregating a small sample of local instances may not be representative of the full picture. Thus while it is intuitive to aggregate local explanations to arrive at a global explanation, without specifying a constraint it is almost impossible to arrive at a complete global picture. However, there exist various ways to define the scope of global explainability in such a way that it is still feasible for XAI methods to approximate the behaviour of a complex model. In this thesis, we build an XAI framework that is **model-agnostic** and generates explanations in a **post-hoc** manner and the scope of the framework is **global**. We achieve the global scope by constraining the model behaviour to a specific dataset.

So far we discussed how different XAI frameworks achieve their goals, Now, let us take a look at one of the most important aspects of any XAI framework - the target audience. Since an XAI framework will be utilised to explain the model outcome to an end-user, human interpretability is considered to be one of the most important aspects. However, the “human-interpretability” aspect of an XAI framework can sometimes be ambiguously defined, especially when it comes to explaining the behaviour of a computer vision model. While many prominent frameworks (*e.g*. LIME (Ribeiro et al., 2016)) use indicative means to present explanation - *i.e*. just highlighting salient regions in an image. While this can still rate high on the human interpretability scale, the actual outcome may be opaque or correlational. This is because a cluster of pixels can conduce different information to different people.

Nevertheless, based on these two distinctions Murdoch *et al*. (Murdoch et al., 2019) defined explainable machine learning as “Extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model“. This definition however leaves out an important aspect of explainability defined by Miller (Miller, 2019), which brings forth the human aspect. According to Miller “interpretability is the degree to which a *human* can understand the cause of a decision“. This decision is further reinforced by (Kim et al., 2016) - “ Interpretability is the degree to which a **human** can consistently predict the model’s result“.

From the above definitions we can see that there is an agreement that human interpretability should be a constant feature in every XAI framework. Especially since its outcomes can have a major effect - practitioners may utilise XAI systems as a debugging tool to uncover bugs in an ML model, especially in a sensitive domain such as healthcare, criminal justice etc. Uncovered bugs, biased and unfair outcomes due to uninterpretable or ambiguity can lead to harmful consequences (McGregor, 2021). Despite this many of the XAI frameworks aim to establish human interpretability with the help of automated systems. Which are then evaluated by human users (Doshi-Velez & Kim, 2017). This is a familiar process within the domain of machine learning and leads to automated systems failing to learn high-level semantics of the data they process generate Jo and Bengio, 2017. This led to practitioners switching to a different strategy called Human-in-the-loop process or HITL for short, that to some extent alleviates the shortcomings of automated systems.

2.1.3. Human-in-the-loop Process

AI systems deployed in real-world systems led to concerns about reproducibility, explainability and controllability (Thiele et al., 2016). To tackle these challenges a new domain of research was proposed called the human-in-the-loop machine learning process (HITL) (Wu et al., 2021). Initially, it was introduced as a means to improve the performance of the automated systems (Zagalsky et al., 2021) but

since then researchers have discovered many ways of integrating human knowledge into their frameworks. The philosophy behind this is very simple. On one hand, humans are ingenious and efficient at many basic tasks (*e.g.* pattern recognition, contextual understanding, seamless translation of low-level to high-level semantics etc.) that even state-of-the-art ML algorithms fail to comprehend. On the other hand, the human brain has limitations when it comes to processing a large amount of information, which computing systems excel. Hence, a framework that combines these two things can potentially be able to overcome the shortcomings of individual elements. HITL processes come under the domain of crowd computing, where the idea is to utilise human knowledge to build robust and responsible systems. However, HITL processes suffer from various problems - and one of the most prominent problems is related to bias. While utilising the power of the crowd for computing is an effective solution, it brings upon caveats of subjective perception, selection bias, sampling bias etc. Thus to avoid these pitfalls there is a need for another paradigm that can complement the HITL process concealing its pitfalls.

2.1.4. Causal Inference

Causal inference is the "discipline that considers the assumptions, study designs, and estimation strategies that allow researchers to draw causal conclusions based on data" (Hill & Stuart, 2015). Drawing causal conclusions is important as it provides clarity to our understanding of how a system works. This is very different from establishing statistical dependence/independence, as the latter may be susceptible to confounding paradoxes or are simply misleading. For example, a barometer reading can be statistically correlated with chances of rain but the reading itself does not cause the rain to fall directly. Confounding mechanisms like air pressure causes rain to fall which in turn also affects barometer reading. Thus, only looking at the barometer reading may give us an indication of rain but is not sufficient to generate an explanation for rainfall. Similar situations can be considered when it comes to explainable AI frameworks, especially for frameworks that aim to explain the learning of computer vision models. While highlighting salient pixels one may conclude that the model is making its decision based on the object that the pixels represent. However, it is plausible that the model is looking at the colour of the object or its shape rather than the object itself. For example, the model may not have any understanding of the high-level concept "bed", but during the training phase, it has noticed that the colour white is strongly correlated with the class label bathroom. Thus upon presenting it with a white coloured bed it predicts a bathroom instead of a bedroom. Any XAI framework presenting the concept of bed as an explanation for a bathroom (or bedroom) may be misleading as clearly, the colour is the confounding factor.

2.1.5. Causal Graphs

So far we have introduced causal inference conventionally. But before we explore more about causality we must get ourselves familiar with some of the formal terminologies associated with causal inference. Any process has an underlying data generating process. This is not just limited to the data itself but the story behind capturing the data. To uncover this process we need a hypothesis. This is where **Causal Graphs** is important. Causal graphs are directed acyclic graphs (DAG) that represent a graphical model of underlying causality.

One of the main motivation behind using such a graphical model is the occurrence of **Simpson's paradox**. When an association is established on population level gets reversed or ceases to exist when the same is divided into subpopulations it is called (Sprenger & Weinberger, 2021). The original example of this comes from (Simpson, 1951). Where he shows that the efficacy of a drug is less effective when considering the general population however the trend reverses when one considers the male and female sub-populations separately. The baffling fact appears to concur that if one to not know the gender then the drug is less effective than when the gender is known. These kinds of examples keep occurring in a lot of social studies but can also be found in various data science practices. According to (Pearl, 2016) the paradox appears because "the story behind the data is missing". And rightly so, given the account, it is impossible to arrive at any rational conclusion. However, if we are to look at the story behind the data - like perhaps the presence of an additional factor independent of the effect of drugs reduces the likelihood of recovery and if that factor is present in most samples when sampling the population data, this can explain the drop is efficacy. For telling such stories graphical models are great tools.

Causal graphs are also very useful to deflate the credibility of seemingly interpretable models (*e.g.*

linear regression). In many situations, XAI research aims to build these so-called interpretable models out of observational data to exhume model behaviour and generate explanations. Many of these models may aim to paint a causal picture but they can be misleading. This is especially true for building a multiple linear regression model without much consideration for the story behind the data. Often termed as the "Table 2 Fallacy" (Westreich & Greenland, 2013) occurs when multiple adjusted effects are estimated from a single model in a single table. Despite being purely correlational, this inadvertently paints a causal picture due to human nature to draw causal conclusions from interpretable/predictive models. A very famous example of this comes from a paper published in Nature that aimed to stratify risk factors of COVID-19 by presenting an interpretable model of the same (Williamson et al., 2020). A bizarre fact emerged - regular smokers were found to be safe from COVID-19. The distorted association occurred due to conditioning on a descendent (mediator or child) of a treatment variable (in this case smoking) which induced a phenomenon called collider bias (Griffith et al., 2020). Thus, even though their predictive model was interpretable it was meaningless (Goren et al., 2020).

2.1.6. Causal Inference terminologies

But before we discuss more causal graphs let us first introduce some terminologies associated with the causal inference paradigm as a whole. First, we clarify the difference between conditioning and intervening as these two are very common terminologies used whenever we interact with a variable. **Intervention** is when we fix a value for a variable, for example, if we open up a barometer and manually adjust the pointer to a specific value, that will be defined as an intervention. On the contrary **conditioning** is when we do not change anything. So for example, if we condition the barometer reading, we leave it as it is. To further simplify it, conditioning can be thought of as filtering - According (Pearl, 2016) *"When we condition on a variable, we change nothing; we merely narrow our focus to the subset of cases in which the variable takes the value we are interested in. What changes, then, is our perception about the world, not the world itself"*

Now within a causal graph exists 3 primary relationships between namely: Confounders(Z), Mediators(M), and Colliders(C). A **Confounder** e.g., **Z**, is that factor which has an effect on other variables, e.g., **X** and **Y**, such that **X** and **Y** show correlation despite not being causally related. A Confounder can be visualised as $X \leftarrow Z \rightarrow Y$. Confounders need to be accounted for when studying the relationship between **X** and **Y**. On the other hand, a **Mediator** is an additional variable **M**, causally related to an independent variable **X** causing an *indirect* effect on the outcome **Y**. A Mediator can be visualised as $X \rightarrow M \rightarrow Y$. Finally, **Colliders** are factors that have a common outcome. Given two independent variables **X** and **Y**, a Collider can be represented as $X \rightarrow C \leftarrow Y$.

Backdoor Criterion The above three terminologies (confounder, mediator and collider) are associated with the "backdoor criterion" (Pearl, 2009). The main idea behind establishing a backdoor criterion is to ensure no form of non-causal association flows when we are trying to establish the causal estimate between a random variable, X and the outcome variable, Y. Only the directed edge between the X and Y is considered to be the causal path and every other edge that perhaps passes through other nodes are considered to be non-causal. Blocking these non-causal paths or backdoor paths ensures that the causal association or causation flows only through the direct edge that connects X and Y. A set of variables Z satisfies the backdoor criterion if -

1. Z blocks all backdoor paths (Figure 2.1) - For this to happen two different cases needs to be considered based on if Z is a collider or not - 1) If Z is not a collider, then conditioning on Z blocks the backdoor path 2) If Z is a collider then it should not be conditioned on to block the backdoor path
2. Z does not contain any descendent of the variable, X.

Average Causal Effect Lastly we briefly introduce the concept of average causal effect (ACE) or average treatment effect (ATE). While graphical models allow us to tell the story behind the data they do not necessarily give the strengths of a causal association between the variables. That is where ACE or ATE are used. The quantify the strength of the causal relationship by taking the average difference between if the (binary) treatment had been administered and if it had not been administered. We will discuss this in detail and present its mathematical formulation when we introduce our framework in the next chapter.

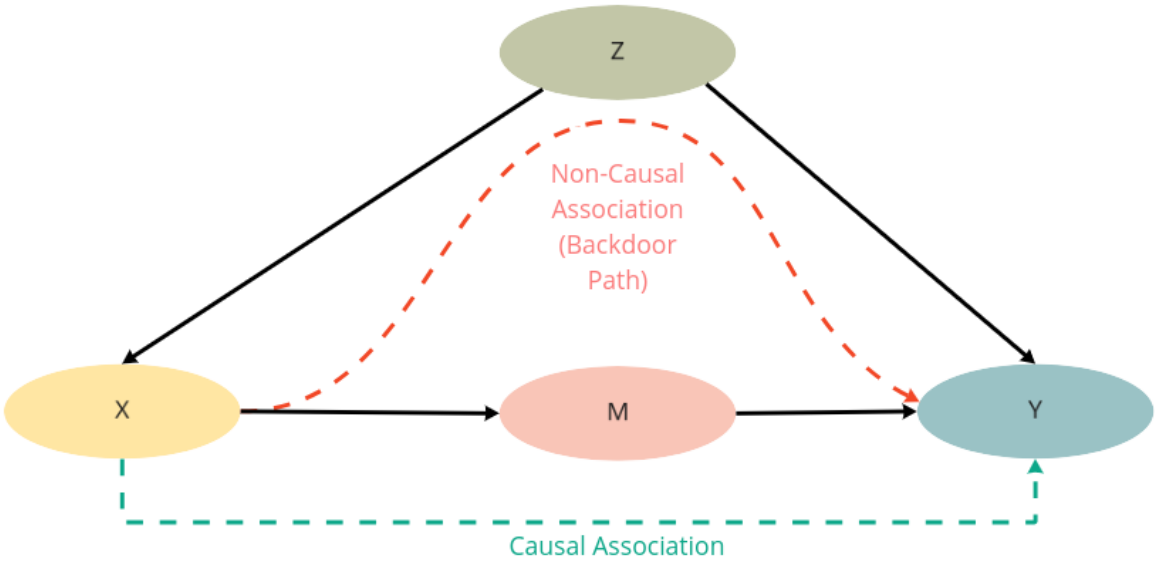


Figure 2.1: Causal graph and showcasing the flow of causal association and backdoor path - here the backdoor paths are those paths that have edges that go into X and also have a path connecting X to Y

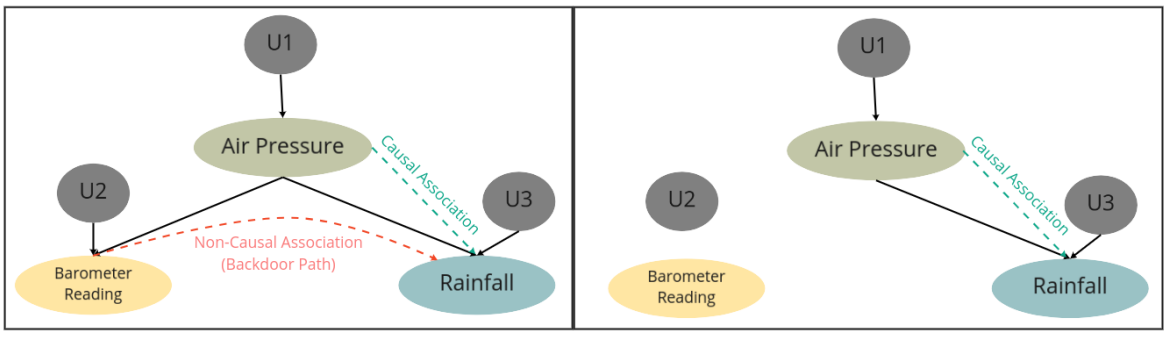


Figure 2.2: Barometer Example: Effect of intervention on a causal graph

2.1.7. Intervening on Causal Graph

Let us now continue with our barometer example again. A causal graph of the underlying problem is presented in Figure 2.2. U1, U2 and U3 represent the unobserved variables that affect the outcomes, these are coagulation of different factors that impact the actual values (*e.g.* air temperature, amount of dust particles in the air, humidity). By utilising this causal graph we can visualise the story behind the data and notice how due to the presence of a backdoor path non-causal association flows from barometer reading to the chances of rain. However, when we intervene in the reading of the barometer *i.e.* to change the reading of the barometer manually we remove all incoming edges to the barometer node. We remove all incoming edges because by performing intervention (changing the reading of the barometer manually) we ensured that the true cause of the barometer reading is our manual intervention and not air pressure or U2. Consequently, the backdoor path is broken as now there exists no edge (non-directional) that connects the barometer with the outcome rainfall leaving only the causal association between air pressure and rainfall. By logically intervening in our underlying graphical model we were able to establish the true cause of rainfall without performing any quantitative analysis.

2.1.8. Causal Discovery

Structuring a causal graph is usually done by experts, or based on prior studies on the same topic. Modelling the relevant factors, mediators, confounders, colliders, and how these are related to each other is not a trivial task. However, the causal discovery process can help ease building causal graphs by inferring the causal structure from observational data. There exist multiple algorithms implementing such discovery processes, each with different assumptions regarding both causal and sampling processes underlying observational data. Glymour et al., 2019 provided a categorisation for graphical methods for causal discovery; here we report only the main ones. Constraint-based causal discovery algorithms, like Peter-Clark (PC) and Fast Causal Inference (FCI) (Spirtes et al., 2000), are based on a complete and undirected graph including all the variables involved and use statistical (conditional) independence tests to prune the edges. On the other hand, score-based models like Greedy Equivalence Score (GES) (Chickering, 2002) start with an empty graph and add edges as long as the scoring function (e.g., Bayesian Information Criterion) increases. Edges are then queried to understand if any removal would further increase the score. Finally, Functional Causal Models attempt to capture the asymmetry that is present between causes and effects by modeling effects Y as a function of causes X, noise ϵ, and some unobserved factors θ.$Y = f(X, \epsilon, \theta)$. Besides graphical approaches to causal discovery, there exist many pairwise approaches that aim to define causal relations between any two variables by evaluating the fitness of the data to an additive noise model (Hoyer et al., 2008), by bidirectionally comparing the standard deviation of the rescaled values of one variable to the other one in the pair (Fonollosa, 2016), or by leveraging asymmetries (Daniusis et al., 2012).

Causal discovery is a powerful tool as traditional ways (i.e., randomised control trials) of uncovering causal relations may be expensive, time-consuming, or impossible. Despite this, their application is not simple and there are many challenges: they might not lead to unique solutions, causal directions might be missing, and faithfulness (i.e., variables connected in the causal graph are probabilistically dependent (Weinberger, 2018)) is sometimes assumed. If not, additional assumptions need to be included (Hyvärinen & Pajunen, 1999; K. Zhang et al., 2015).

2.2. Related Work

2.2.1. Explanations in Philosophy

On the topic of explanations in the field of XAI, Miller's survey (Miller, 2019) was one of the first studies mentioning causality as a possible means to implement XAI frameworks and tackle the limitations of existing methodologies. Particularly, Miller points to the *Ladder of Causation* by (Pearl & Mackenzie, 2018) in which explanatory questions are grouped in three classes: what-questions (e.g., "*What event happened?*"), how-questions (e.g., "*How did that event happen?*"), and why-questions (e.g., "*Why did event that happen?*"). Along those lines, (Buijsman, 2022) reports the properties a *good* explanation should have: first, a rule answering *why* we got a specific output, and second a counterfactual component aimed at answering *why X occurred rather than Y*. Furthermore, Buijsman also conceptualised the depth of an explanation in terms of abstractness of variables and generality. Having a more abstract explanation allows us to answer more why-questions, but this needs to be balanced with the specificity of the explanation (i.e., the information should be relevant to model outcomes). On the other hand,

generality is related to the number of inputs covered (i.e., breadth), balanced against the correctness of the explanation on those inputs (i.e., accuracy). The former refers to the preference for more abstract variables, that is, in his words

> "As long as both are correct, we tend to see the explanation using *red* as better."

when given the following explanations:

- The pigeon pecked because it was presented with a scarlet stimulus
- The pigeon pecked because it was presented with a red stimulus

Having more abstract explanations allows us to answer more why-questions. However, increasing the level of abstractness means lowering the specificity of the explanation. To be considered *good*, an explanation should be abstract enough by incorporating only relevant information, i.e., not compromising on specificity. On a related note, generality is tied to abstractness and can be quantified in terms of breadth and accuracy. The former simply means that an explanation can cover more inputs. The latter means that the explanation is correct for every input in its range. Similar to abstractness and specificity, the relation between breadth and accuracy is one of balance. Furthermore, they also highlighted the relevant aspect and structure of an explanation. For the most part, past works in philosophy of science and social sciences are critical of XAI given a large number of definitions, their sparsity, and lack of clarity across the literature. But in general, various aspects of what makes an explanation as the concept of explanation is undeniably defined in better terms outside of the domain of computer science. We take inspiration from these discussions and ground our proposed method on the result of these works.

Differently from existing approaches, in our work, we specifically take an interventionist account (Grimsley et al., 2020) for generating explanations by leveraging causal inference methods on top of crowd-generated concepts (discussed in the remainder of this section).

2.2.2. Causality in Explainable AI

There have been various attempts at implementing the concept of causality into the field of XAI, by drawing inspiration from the Causal Inference field, especially via generating counterfactual-based explanations. Works specifically related to Causal Inference will be presented in more detail later on, in the Background section. As a reference point, counterfactual examples differ from adversarial ones as the former aim to define changes in the input so that alternative outcomes happen instead of the original one (Brughmans et al., 2021), the adversarial examples are meant to fool the attacked model and make it fail in its task (Freiesleben, 2021). Counterfactual explanations can be obtained by altering the values assumed by the different variables governing the given phenomenon through interventions. Interventions are not new in XAI frameworks but, to produce meaningful results, they must be designed carefully (Woodward, 2003) so that they precisely target variables of interest. Several approaches have been proposed to generate counterfactuals through heuristic searches, instance-based strategies, decision trees, or by framing optimisation problems. Guidotti, 2022 provides a thorough review of these approaches. Two examples are the ones by Wachter et al., 2017 and Dandl et al., 2020, both of which are based on minimising loss functions that constrain certain desired properties (e.g., the high similarity between the actual instance and the counterfactual). Counterfactuals have also been used in the NLP field. For example, scholars have created counterfactually-augmented datasets that enabled them to produce models which learn causal features and achieve better performance on unseen data (Kaushik et al., 2019; Kaushik et al., 2020). More specific to computer vision, Goyal et al., 2019 proposed an approach that, given two images, identifies the key discriminative regions in them such that swapping those regions leads the model to change its prediction. The approach is specific to convolutional neural networks as the authors focus on the feature extracted in the earlier layers of the network. Besides the plethora of approaches proposed to generate counterfactual generation, Guidotti, 2022 raises an important point by uncovering, based on existing counterfactual explainers, how researchers have mostly overlooked causality thus far. To the best of our knowledge, ours is the first approach focusing on this dimension of counterfactual explanations in the field of XAI.

2.2.3. Explainability of Computer Vision models

In the context of computer vision explainability, saliency is the most widely applied approach. Saliency is a local, post-hoc explainability method that highlights the most important pixels in a single image

with respect to the model prediction (Simonyan et al., 2013). Saliency can be computed by computing the gradient of the activation functions (Selvaraju et al., 2019; Simonyan et al., 2013), by backtracking the features to the inputs (Shrikumar et al., 2017; Bach et al., 2015), or with more sophisticated approaches like SmoothGRAD (Smilkov et al., 2017). From a different angle, Kim et al., 2017 provides a concept-based approach to explaining CV models by introducing the notion of Testing with Concept Activation Vector (TCAV) and using it to perform translations between the internal states of a model to human-friendly concepts. Ghorbani et al., 2019 later expanded on TCAV by identifying concept-level information across different images, clustering them, and testing their importance. The main disadvantage of these approaches is that the highlighted regions still need interpretation. Finally, two more recent approaches by Balayn, Soilis, et al., 2021 and Sharifi Noorian et al., 2022 use crowdsourcing to address two XAI problems: concept extraction for global model interpretability and unknown unknowns characterisation respectively. Considering the existing contributions in establishing procedures to answer the *why* aspect of explanations, our study complements those by adding a counterfactual analysis. We do so by eliciting people's cognitive abilities to collect human-understandable concepts as hypotheses to be further validated through causal inference. We focus on analysing the causal effects different concepts in images have on the final model prediction. By taking a causal stance in explaining model behaviour, we are enabled to consider confounding factors as well as perform interventions on individual concepts to explain a model's output.

3

Framework

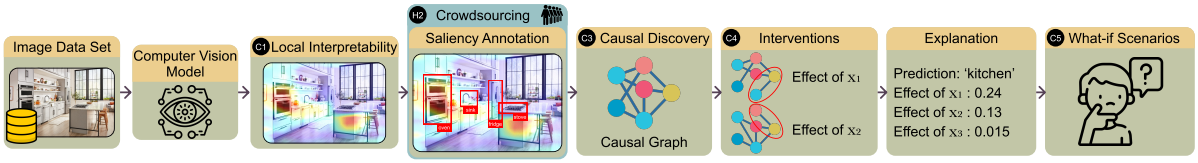


Figure 3.1: Overview of the CHIME workflow.

In this section, we discuss the CHIME framework and the underlying motivations. CHIME is an ensemble of different methods applied toward the common goal of identifying and explaining the behaviour of Deep Learning models for Computer Vision, given their predictions on a set of images. The grounding for this is derived from interventionist account of philosophy (Grimsley et al., 2020; Buijsman, 2022; Woodward, 2003; Miller, 2019) which is then combined with logical structure of causation proposed by Pearl *et al.* (Pearl & Mackenzie, 2018). The interventionist account states that if X (estimand) causes Y (outcome) then any intervention on X, represented by do(X) := $X_{initial} \rightarrow X_{final}$ where $X_{initial} \neq X_{final}$ will generate the subsequent change the value of Y ($Y_{intial} \rightarrow Y_{final}$) such that $Y_{initial} \neq Y_{final}$. The idea behind intervention was already depicted in Section 2.1.6 with the help of a graphical model in Figure 2.2. Subsequently, we now apply the counterfactual account to further solidify the outcomes of the framework. This is also reminiscent of the three layers in causal hierarchy presented by Pearl (Pearl & Mackenzie, 2018). The first level is **association** - this is our initial hypothesis about what the model learns; the second level is **Intervention** - Establishes a causal perspective by intervening on learned features in the first level, and finally, the third level is **Counterfactuals** - helps the framework generate a more extensive set of answers, especially for "what-if" questions.

Given this high-level overview of the framework, fully visualised in Figure 3.1, we will explain each more in detail each component in the remainder of this section.

3.1. Data Collection - Establishing Correlation

The first component of our framework aims to capture the correlation between model prediction and corresponding salient pixels. To do this we first need two inputs - 1) A computer vision model trained for the image classification task and 2) A dataset - whose instance will be used to explain model behaviour.

3.1.1. C1: Saliency Map Extraction

Once we have the pre-trained model we can use it to make predictions on all instances within the dataset provided. Since the datasets are generally large, it can be costly to consider all instances, thus we perform random normal sampling across different classes within the dataset to select a small subset of images. Once the model has made predictions, we then utilise that information to capture salient pixels. This is generally done using saliency map detection algorithms as described in Section 2.1.6. While there exist many variants of generating saliency maps we utilise a method called SmoothGrad (Smilkov et al., 2017). SmoothGrad is an architecture agnostic saliency map generation technique that

generates sensitivity maps by adding noise to the underlying image. This algorithm to some extent also follows the interventionist account (adding noise to the image can be considered as intervening on the image) thus aligning well with our causal framework. Upon identifying the sensitivity maps which are nothing but heat maps depicting most activated pixels with respect to the model predicted class, we can begin translating the blob of pixels to meaningful high-level semantics.

3.1.2. H2: Human Annotation of Saliency Maps

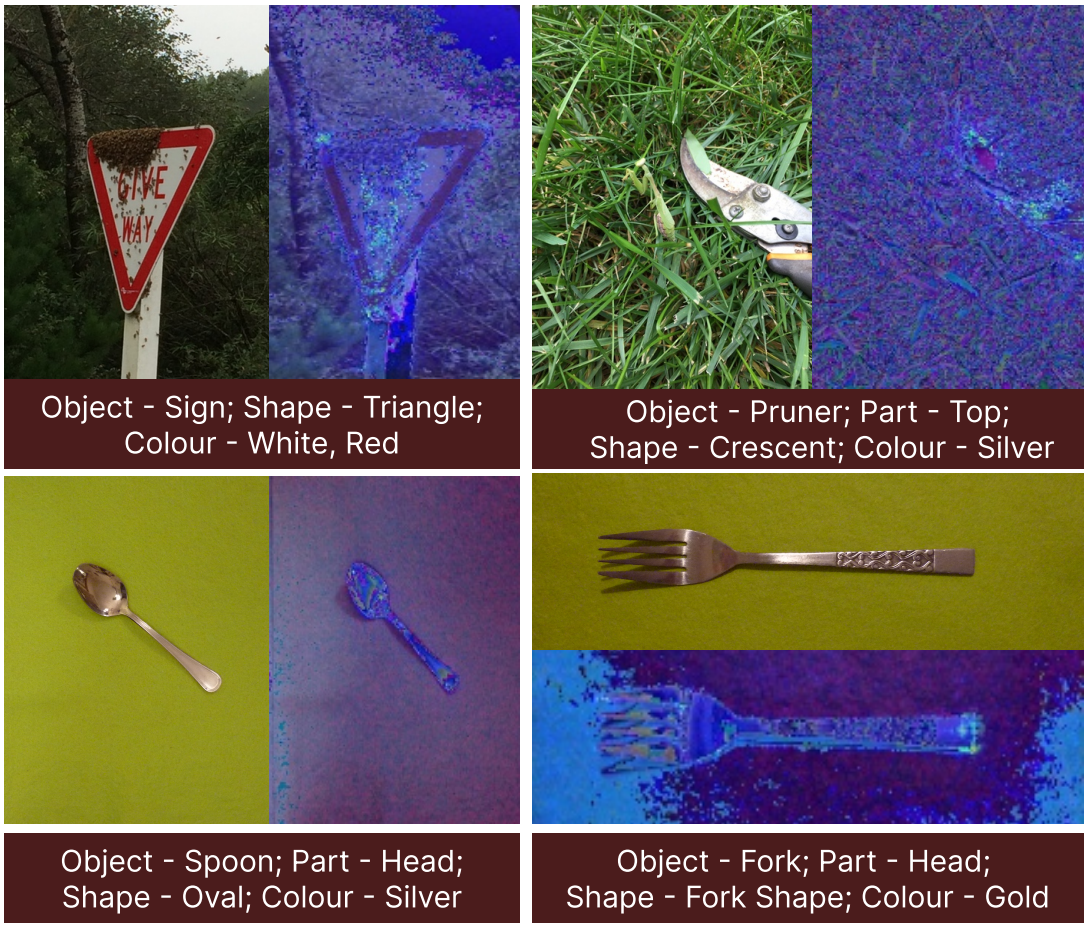


Figure 3.2: Examples of annotated concepts from different biased models: Neg. Set Biased Inception (top-left), Neg. Set Biased Squeezenet (top-right), Colour-Biased Inception (bottom-left), Colour-Biased Inception (bottom-right)

As mentioned earlier a cluster of pixels represented by the heat map albeit interpretable can be ambiguous - they can refer to different concepts: an object, its shape, or its colour. There is no straightforward way to distinguish these individual concepts. Automatic methods such as object detection (Lin et al., 2014) can be utilised but despite their high accuracy these methods are generally very much limited by the supervised labels they were trained on. This can affect the generalisability of the framework. Furthermore, doing so would introduce a black-box component in our framework - and explaining its behaviour will add additional overhead. Thus, motivated by (Balayn, Soilis, et al., 2021) we apply a crowd computing approach. The idea here is to involve crowd workers and elicit their cognitive abilities to annotate salient patches. This is fairly effective as human cognition is highly accurate at detecting concepts and it also provides us with a wide variety of meaningful concepts. Our approach differs from Balayn *et al*. in terms of the crowdsourcing task design. In the original work for each saliency map entity attribute, pair was captured. However, based on our initial hypothesis (if the model is looking at the object, its colour or perhaps its shape), we needed more specific labels. Thus our task design incorporated fields that specifically captured the name of the 1) Concept (or object) 2) the colour of Concept 3) the shape of the Concept. Additionally, we capture the part-of relationship - sometimes,

the heat-map may not identify the concept as a whole but rather part of it *e.g.* leg of a chair or head of a person etc. By implementing this framework implicitly we were also able to formulate a template for causal discovery and validate our hypothesis. Apart from this, the task design also incorporated a drop-down suggestion field to direct/suggest annotators toward the intended values. These were not restrictive, meaning annotators could still select values from outside of the suggestions provided. This gave a good balance between indicative and free-form design. The template used to capture annotation data from crowd workers is shown in Figure 3.3

Once the concepts are collected they were checked for spelling violations and consequently corrected; spaces and other special characters were removed the cases were lowered. The rest of the pre-processing was done manually. Specifically, colours were made more general *i.e.* if the colour was annotated as *light green*, it was replaced with *green*. This was done to ensure that the concepts that needed to be explained had a higher level of abstractness aligning our work with the hypothesis presented in the work of (Buijsman, 2022). Note that there still exist cases where certain colours such as the colour olive - which is a combination of yellow and green were left untouched as the current version of the framework does not have the capabilities to detect and resolve such conflicts heuristically. However, the colours were annotated as “yellowgreen” were replaced with “yellow” based on the precedence of yellow over green, similarly “greenyellow” were replaced by the colour green. Apart from this the pre-processing step also removed noisy and meaningless annotation. For example, some of the annotations that considered “background” as a shape were removed. For a small number of instances, the annotations were wordy. For example, in one instance one annotation for colour read ‘transparent, but having a reflection‘, these instances were manually verified and fixed. While these could also have been removed but as the captured data was already limited, we aimed at keeping as many instances as we can. Some examples of annotations after pre-processing for different models are shown in Figure 3.2.

While we present a very specific design based on our hypothesis, in practice any form of crowdsourcing task design can be implemented depending upon the hypothesis of the stakeholder.

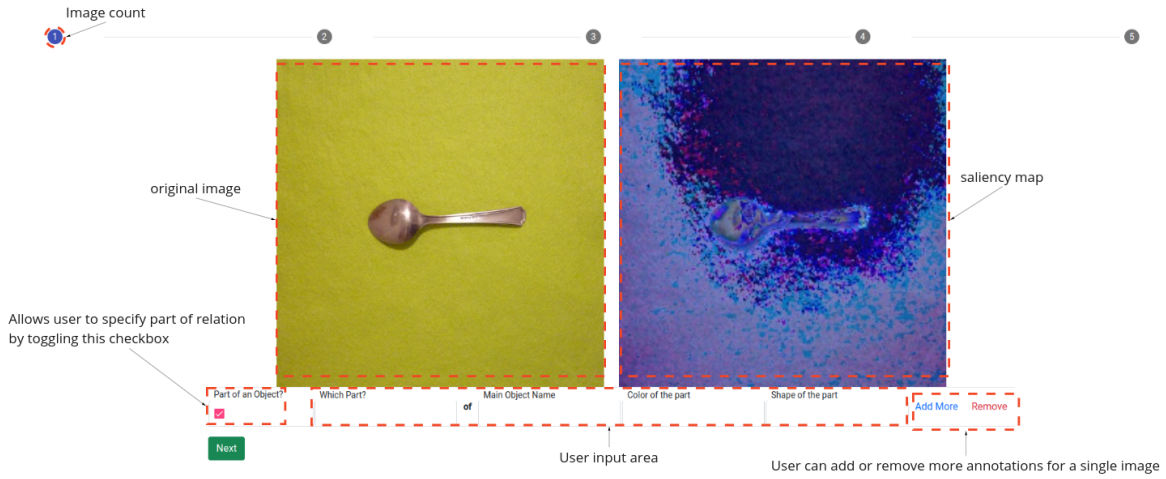


Figure 3.3: Crowdsourcing task design

3.1.3. Concept Aggregation

Acquired concepts in the previous steps are local *i.e.* they belong specifically to individual image instances. TO arrive at a global picture we need to aggregate the annotations. Now, in our background reading, we mentioned that aggregating local interpretations can be incorrect, however, in this scenario, we have a constraint in the form of the dataset *i.e.* our framework limits its explainability prowess to the dataset provided as input. With this limiting scope, this aggregation step can be an effective tool for translating the local scope into a global one. We aggregate the individual concepts on a class level *i.e.* we combine all the annotated instances belonging to the same class. However, the observational data is still correlative in nature. We have collected data points which we believe can be correlated with

model prediction. And as already established in previous sections interpretability does not come only from fitting data to a simpler model but rather it is a combination of understanding the structure of the data(causal graph) and then building a model based on it. Thus we need a way to formulate and then validate the observational data collected - this is where we introduce the causal discovery component of our framework.

3.2. C3: Causal Discovery

To understand the underlying structure of the annotations we collected we build causal graphs for each class. This also helps us to exhume the presence of confounders. We construct our causal graph in two ways - 1) **Template-based** - Causal graph generated by this method requires manual intervention and is primarily based on a stakeholder's initial hypothesis or domain knowledge. This initial hypothesis can be formulated in several ways - from an XAI practitioner aiming to uncover a specific trait of a model (*e.g.* effect of a particular colour on model prediction) to a model creator debugging their creation for bugs (checking if the model prediction is sensitive to out of domain concepts to prevent adversarial attacks). 2) **Automated** - This way of generating causal graphs requires no prior knowledge but they require the usage of causal discovery techniques that generate causal graphs from observational data. As part of this thesis, we employ a pairwise causal discovery algorithm called Conditional Distribution Similarity Statistic (CDS). But again, it can be left at the hands of the stakeholder who can select their preferred automated causal discovery algorithm and use it as a plug-and-play component with our framework.

Template-based Causal Graph Creation Building causal graphs is not trivial and may require domain-specific knowledge to be effective. Despite this, a template-based causal discovery can allow a practitioner to construct and validate their hypothesis about certain behaviour of the model. We have touched upon this briefly while discussing the crowdsourcing task design. Based on the requirements of stakeholders one can alter the design of this template to incorporate their hypothesis. We hypothesised that perhaps the model is not looking at the object itself but rather part of the object or its colour or perhaps its shape. Thus we similarly constructed our causal graph template - we considered the primary object as estimand and its colour or shape as the mediator. The model prediction is considered to be the outcome variable. The collected data for each of these 3 nodes were filled in their respective place *e.g.* Bed (primary object) was placed as the estimand, black (colour) was considered to be the mediator and bedroom (class label) was input as the outcome variable. Table 3.1 shows the templates that were considered.

Templates	
object/part-of object → colour	colour → label
object/part-of object → shape	shape → label
object → label	

Table 3.1: Templates from template-based causal discovery

The rationale behind this is fairly straightforward: the presence of an object may directly affect the prediction label, but at the same time, it causes the presence of a certain colour in the image, and objects define shapes, both of which can affect the model outcome as well. This kind of simplistic reasoning avoids the caveats of Occam's razor succinctness principle (Fonollosa, 2016). Figure 3.4 depicts a simplified example of a causal graph generated using a template-based discovery method.

But due to the modular nature of this framework, one can just as well inject their design and validate their hypothesis about the model's behaviour. This however establishes a tight coupling between the template-based causal discovery and crowdsourcing task design.

Pairwise Causal Discovery As previously discussed in the Background section, causal discovery can alleviate the process of building a causal graph by discovering causal structures from observational data. The template-based causal graph generation requires prior knowledge and an initial hypothesis, in that sense it is dependent on an input given to the framework. However, this type of supervision may not always be available. When a stakeholder has limited knowledge about the domain or perhaps the

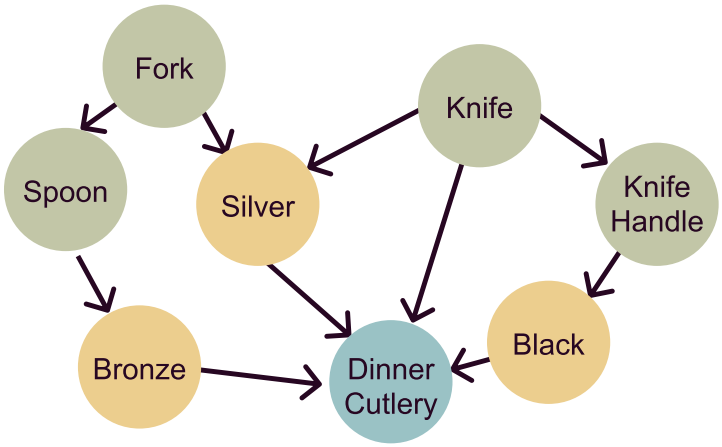


Figure 3.4: Template-based causal graph example, including primary (green) and mediating (yellow) concepts, and the model outcome (blue) for the "Dinner Cutlery" class.

domain itself consists of many estimands that have intricate relationships amongst them than creating a template may be time consuming and complicated task. This is where automated causal discovery methods are useful. In our scenario, we utilise the Conditional Distribution Similarity Statistic (CDS) algorithm by (Fonollosa, 2016), given the discrete nature of the crowd-powered annotations. The CDS method works on a very simple assumption - "the shape of the conditional distribution $P(Y|X = x)$ tends to be very similar for different values of x if the random variable X is the cause of Y". In the original paper, there exists different features measuring these distributional differences and were then input into a Gradient Boosting Classifier to predict the direction of causality. For our implementation, we considered the Causal Discovery Toolbox implementation of this algorithm (Kalainathan & Goudet, 2019) that implements only one of the features mentioned in the paper. Specifically, it focuses on the standard deviation of the conditional distribution.

Since it is a pairwise method it returns the direction between two variables. Thus all collected concepts and their corresponding colour, and shapes were considered to be nodes in a non-directional complete graph. Then the CDS method was applied to each pair of nodes in the graph to determine their direction of causality. Mathematically, for all pair of nodes (X_a, X_b) where $i \neq j$ If X_a is a cause of X_b then for all values $x_1, x_2, ..., x_n \in X_i$ the standard deviation of the conditional probability distribution $P(X_a|X_b)$ will be lower than that of the reverse relation *i.e.* $P(X_b|X_a)$. Before calculating the standard deviation both distributions are standardized to have a mean of 0.

$$CDS(X_a, X_b) = D(X_b, X_a) - D(X_a, X_b) \tag{3.1}$$

where,

$$D(X, Y) = \sqrt{\frac{1}{M} \sum_{i=0}^{M-1} var_{X_i}(P_n(Y_i|X_i))} \tag{3.2}$$

In Equation 3.2 P_n represents the normalized conditional probability and var_{X_i} represents variance over X_i. The CDS score returns a value of 1 or -1 specifying whether the causal direction is between $X_a \rightarrow X_b$ or $X_b \rightarrow X_a$ respectively. If no causal direction can be defined then it returns 0.

Based on requirements one can plug in any form of causal discovery algorithm or supervised template based on expert knowledge. This increases the versatility of our framework and allows stakeholders to experiment with different techniques to achieve their desired goals.

3.3. C4: Performing Interventions to Determine Causal Effects

By generating the causal graph, we have an overview of the hypothetical model behaviour. However, by itself, the graph does not provide any information regarding the causal strengths of individual concepts with respect to the model outcome. These causal effects can be estimated through intervening

on the estimand. The causal effect values enable us to rank concepts based on the magnitude of their effects on the model outcome. Practically, we conduct linear regression on the crowd-sourced concepts and then observe the changes in the output based on carefully performed perturbations (i.e., interventions) to its inputs. As already mentioned interventions can be formulated as $P(Y|do(X), Z)$, where X represents a single concept, $do(X)$ is the action of setting the variable X to a particular value, and Z is the set of confounders on which the estimates are conditioned on, to not obtain distorted associations with the model output. We perform interventions on causal graphs by removing all incoming edges to a particular node, thereby removing their influence on the intervened variable, allowing us to capture the direct effect a single variable X has on the outcome Y. For our use case, we define treatment values as 1 and control values as 0 representing the presence or absence of the concept respectively. While we intervene on every available node except the outcome node certain criteria are considered while selecting the set of conditioning variables (Z). This is done to ensure no unwarranted association flows into our underlying estimand (X) and the outcome variable (Y) failure to do this may result in erroneous. To tackle this we need to establish that Z is a part of a sufficient adjustment set. A set of variables is considered to be in *sufficient adjustment set* if it satisfies the backdoor criterion (Neal, 2020). Once a variable is identified to belong to the sufficient adjustment set then only we can control/condition on it to estimate the causal effect. We utilise a process called Conditional Outcome Modelling (COM) to calculate the average causal effect (ACE). The formula for ACE is -

$$ACE = \mathbb{E}_{\mathbb{Z}}[\mathbb{E}[Y|X = 1, Z] - \mathbb{E}[Y|X = 0, Z]] \tag{3.3}$$

Where Y is the outcome variable, E_Z represents the expected value of all values of Z and Z is a variable in a sufficient adjustment set, and as mentioned earlier setting the value of X to 1 and 0 signifies the presence or absence of that concept. To get the conditional distribution one can utilise any machine learning model. But utilising complex black box modelling techniques may add an extra overhead of having to explain the behaviour of that model also. Thus for our project, we utilise linear regression, as it is a simple and interpretable model. It can also be proven that the regression co-efficient represents causal strength if there exists no backdoor path between the estimand and the outcome (Neal, 2020).

Aside from the mathematical formulation, logically this amounts to validating our hypothesis. 1) By satisfying the backdoor criterion we ensure that the flow of information between the estimand and the outcome is unadulterated. 2) Upon intervening upon the estimand we then ensure that our estimand is free of any effect by other unobserved variables and then validate if the absence or presence of the estimand makes a difference in the outcome.

3.4. C5: Answering *what-if* questions

Thus far, we have achieved the second level using (Pearl & Mackenzie, 2018) *i.e.* by intervening on the estimands we have managed to establish a causal association between our intended concepts and the model's outcome. This allowed us to paint a causal picture with regards to what the model has learnt. Not only that but by quantifying the causal strengths we can also see which concepts have a stronger causal relationship as compared to the other. However, there is one more step in the causal hierarchy that is missing. The counterfactual level allows us to generate answers to *what-if* questions. Generally in social sciences, the differences between counterfactuals and interventions are well defined. Take for example a hypothetical study to establish whether smoking causes cancer or not. We can source participants and divide them into two groups, participants from one group are asked to smoke cigarettes (treatment group) and participants in the other group were asked to do the opposite (control group). Smoking here is the intervention we perform to gauge smoking's effect on cancer. Now we let the study pan out and after some years we capture our results. If during that time, the group that smoked had more cases of cancer compared to the control group then we can conclude that smoking is a cause of cancer. However, at this stage one might enquire "what-if" the group that smoked cigarettes did not smoke cigarettes, would that have reduced their chances of cancer? Note the subtle difference, during step 2 of the causal ladder (intervention), we had a world where none of the participants had cancer and we intervened to establish whether smoking causes cancer. Now, as for step 3 of the causal ladder, we know that a group of people have already smoked cigarettes but we wonder what if they hadn't and thereby creating a counterfactual reality. While the distinction is clear in a real-world scenario from the perspective of an image and intervention by itself can be corroborated in a counterfactual scenario. For example, when we intervene in a concept that is present in an image we already go back in time to a

state when the image was being taken and change an aspect of it. This is because a digital image is an event that has already taken place any form of intervention results in contradicting that reality that took place.

3.5. Toy Example

To further clarify this let us consider a toy example that explains how each component in the framework works. Let's consider the case of binary scene classification - using an underlying deep learning model we classify whether an image is "bedroom", or "not bedroom". First, we present the trained model and the entire dataset to the framework. Upon presenting it with n randomly sampled images from a dataset we extract the salient pixels (C1) and then annotate them using crowd workers(H2). Upon aggregating the annotations we notice that the *Primary Concepts* are {bed, table}, and the *Mediating Concepts* are {black, red}.

The findings reported from the annotation stage are reported in Table 3.2.

Image	Primary Concept	Mediating Concept	Model Prediction
1	Bed	Blue	Bedroom
2	Table	Red	¬ Bedroom

Table 3.2: Toy example: Captured annotations

For simplicity, We apply template-based causal discovery for this toy example and hypothesise that both primary concepts are causally related to the model outcome under the influence of the mediating concepts. Based on this we formulate as many as

Given that knowledge, we build linear models with the following structure:

$$\begin{aligned} y &= a \cdot PC + b \cdot MC + \xi_1 \\ MC &= c \cdot PC + \xi_2 \end{aligned} \tag{3.4}$$

where PC and MC represent a primary concept and a mediating concept respectively; ξ_1 and ξ_2 are the noises associated with the underlying linear models. To estimate the values of the coefficients a and b, we construct two separate linear regression models, one to compute the causal strength of object → bedroom, and another for colour → bedroom, where the object is a confounder. In both cases, interventions are performed on the estimand (bed, table, red, blue) to ensure that it is not influenced by any observed or unobserved confounder. When estimating the causal effect of object → bedroom, we do not consider colour as a confounder. This is primarily due to the colour being a descendent of the treatment variable which might induce collider biases (Cinelli et al., 2021).

In both cases, the outcome variable is the model prediction. To further simplify the process we consider binary interventions, i.e., the presence and absence of the said object. Employing interventions, we calculate the causal strengths of those concepts, i.e., the coefficients of the linear model. The higher the coefficient, the higher the causal strength.

The benefits of identifying such a coefficient are two-fold. Firstly, it helps us identify the causal strength and helps us answer counterfactual questions. Secondly, they help us calculate the remaining unknowns of the equation - the errors (ξ). This allows us to construct a parametric model based on which we can answer questions like "What if the images had a red bed? or a black table?" when trying to understand the behaviour of the toy model. Note that this is a direct consequence of our framework's capabilities, however, it is not implemented as part of this project.

4

Experiment Set up

Evaluating an XAI framework can be complex as there exist no well-established benchmark standards that can be used for comparisons (Yalcin et al., 2021). This issue generally stems from different XAI frameworks catering to different system goals (Mohseni et al., 2018). To further complicate the matter there exists no gold standard *i.e.* we can only hypothesise what the model has learnt or make it overfit or sensitive to certain features but even then it is not guaranteed that the model behaves in the way we expect it to. However, there exists metrics that are commonly used in the existing XAI domain. We take motivations from such work and design our experimental setup in such a way that the framework is evaluated both from the XAI and the causal perspectives.

4.1. Fidelity

Amongst standard XAI metrics, **Fidelity** is considered one of the most important properties of an explanation (Molnar, 2022) as it represents the ability of an XAI framework to approximate model behaviour. However, Fidelity is interpreted differently across literature and implemented differently based on the suitability of the framework (Balayn, Lofi, et al., 2021). Z. Yang, 2019 defined fidelity as the model's ability to generate outcomes that closely match that of population distribution. The population distribution is the known distribution. Mathematically he defined Fidelity as

$$F(x, y; H) = 1 - |P_{model}(y|x) - P_{pop}(y|x)| \tag{4.1}$$

where F(x,y; H) is the fidelity of model H given an input sample x and outcome y. $P_{pop}(y|x)$ is the known probability distribution of y given x and $P_{model}(y|x)$ is the model's prediction distribution. In our experiment, we take motivation from this definition and present the fidelity in such a way that it is more interpretable and demonstrative in nature rather than quantitative. Firstly to establish the known probability distribution by injecting biases and fine-tuning the models for a sufficient number of epochs so that their behaviour is skewed toward those biases. Specifically, we inject Sampling bias and Negative Set bias and then utilise our framework to identify those biases in the generated explanations.

Sampling Bias or Selection bias or Collider bias is introduced when the dataset is sampled in such a way that it introduces a spurious correlation. Consider the PASCAL VOC dataset (Everingham et al., 2010) bird and sheep class. Since most pictures of sheep are most likely to be taken with a background containing grasslands and similarly for birds it will be the sky. Despite the original idea to extract features of birds and sheep the model will now focus on the background features as a cue to classify the images into birds and sheep. Deep learning models are known to utilise these spurious correlations to make predictions (Y.-Y. Yang et al., 2022). In our case, we intentionally introduce this bias to make the model behaviour more predictable.

Negative Set Bias This is a very intriguing and also common form of bias that is present in datasets used for image classification. As defined by (Torralba & Efros, 2011), negative set bias in datasets occurs when apart from the positive instances - the instances or concepts that the label represents, there exist many instances of other concepts, called negative instances thereby referred to as the

negative set bias. Every image represents a piece of specific information, and it is rarely the case that a single image will contain only one object corresponding to the label or class it belongs to within a dataset. Take the dog class in imagenet (Deng et al., 2009) for example - apart from the positive instance of a dog it has grass, trees, roads, goldfish, people, snow etc. which all can be categorised as a negative set. Now any computer vision model aiming to classify an image as a dog will take all these visual clues, however, if there are other competing classes (e.g. grass, trees, people) within the same dataset which considers the negative set of dog class then there is a possibility that the model will misclassify. While the positive sets are well defined within a dataset, negative sets are never documented. However, they are important to uncover as they can assist in segregating the intended classification object from other visual phenomena. A detailed description of how these biases were injected will be discussed in Section 4.6.2.

4.2. Accuracy

Accuracy is a widely used metric in XAI-related fields. This is generally due to its quantitative simplicity and fairly transparent conclusive prowess. A different definition is defined for a different context and is often correlated with Fidelity (Molnar, 2022) but since for our framework, we define fidelity demonstratively we present accuracy as a separate quantitative measure. Accuracy in XAI literature is defined as the underlying framework's ability to correctly estimate model prediction with the generated explanation, especially on unseen data. For generating an explanation the model's prediction is the input to the XAI framework but for calculating accuracy the order is reversed. We first source the top 5 concepts ordered by their causal strength to a particular class. While any number of concepts can be considered we empirically choose the number 5 to make the outcomes more compact and the results more presentable. These concepts are then put into the Google image search engine using a python framework [1]. For colours and shapes the prefix "colour" and "shape was used respectively to ensure the images contain the underlying colour and shape as without these prefixes majority of the results were erroneous. For primary concepts, no alteration was done. In total 10 images were collected for each concept resulting in a total of 50 images per class. These images were then input into the model and the corresponding top 2 predictions were captured. If the model's top most predicted outcome matches the class for which causal concepts were used to generate the input images then the Accuracy@1 score inflates and similarly if the class belongs in the top 2 highest predicted outcomes then the Accuracy@2 score inflated.

$$1Accuracy@1 = \frac{\text{Number of Target label in top 1 predicted outcome}}{\text{Total Number of images}} \tag{4.2}$$

$$Accuracy@2 = \frac{\text{Number of Target label in top 2 predicted outcome}}{\text{Total number of images}} \tag{4.3}$$

4.3. Causal Verification

Since our framework relies on discovering causal concepts we need to ensure that the established concepts are in fact causal in nature. Inspired by the idea presented in (Xu et al., 2020), we check if top causal concepts are relevant for a particular class as compared to the weak causal elements (concepts with the bottom-most causal effect score). Like calculating accuracy, for each class, the concepts were ordered based on their causal strength. Then we hypothesise that the top 5 and bottom 5 causal concepts from the ordered list represent strong causal concepts and weak causal concepts respectively w.r.t the underlying class. While other concepts may also be useful in terms of generating explanation and a case can be made to perform a 50-50 split, *e.g.* if there are 20 total concepts for a class, the top 10 can be considered into the strong causal category and the bottom 10 can be considered as the weak causal category, we choose to go with top 5 and bottom 5 concepts as it makes the metrics more specific. Once the strong and weak causal concepts are identified we perform causal verification by evaluating the following inequality.

$$P(\text{effect}|\text{strong cause}) > P(\text{effect}|\text{weak cause}) \tag{4.4}$$

[1] https://github.com/hardikvasa/google-images-download

where considering images with the top-5 causes

$$P(\text{effect}|\text{strong cause}) = \frac{\text{correct predictions}}{\text{\# of images with top-5 causality}} \tag{4.5}$$

and, similarly, considering images with the bottom-5 causes

$$P(\text{effect}|\text{weak cause}) = \frac{\text{correct predictions}}{\text{\# of images with bottom-5 causality}} \tag{4.6}$$

We hypothesise that the model's top prediction should favour the underlying class for the top causal concepts rather than the bottom ones.

4.4. Coherence

A human-centric framework should focus on explanatory metrics that are important from a human's perspective. With that regard, (Thagard, 1989) in his Theory for Explanatory Coherence stated that an "explanatory hypothesis is accepted if it coheres better overall than its competitor". This has later been correlated with how a human makes judgements on explanation (Ranney & Thagard, 1988). While the caveat of utilising coherence as a metric in XAI is that it may not always concur with prior beliefs of a human thereby may lead to an explanation being rejected (Miller, 2019). Nevertheless, assuming that the explanations generated by an XAI framework are independent of a stakeholder's biases regarding perceptual similarity, one can define coherence as the framework's ability to generate similar explanations for similar data instances (Molnar, 2022). Note that the term "similar instances" can refer to many things. For example, it can be considered at a class level, but doing so will require us to summarize a diverse set of complex information. Which can be a complex process itself not to mention lossy conversion. Thus, while other metrics are considered at a global level aligning with our framework's attribute, for coherence we take a different route and focus on local instances. This also showcases the flexibility of our framework.

4.4.1. Finding Similar Instances for Coherence

The first step to establishing coherence we need to establish a way to discover similar instances. We do this by considering the HSV colour model given its strong relation with human perception of colours (Paschos, 2001). Once HSV features are extracted, we apply Isomap to obtain a 2-dimensional representation (embedding) of those features allowing us to visualize and manually validate them. Then we construct a similarity matrix using these 2D embeddings. Finally, the top 10 most similar images are paired using the Manhattan distance.

This is an automated procedure that is based on empirical evidence, and thus not fully accurate. To alleviate this the authors manually validated the similarity of the generated pair by subjectively inspecting 45 subsets of image pairs. Subjective similarity has been used as ground truth for automated similarity techniques (Li et al., 2020).

The similarity evaluation was done on two metrics - 1) based on colour similarity and 2) based on object similarity. In terms of colour similarity 34/45 = 75% accuracy was achieved but 16/45 = 35% accuracy was achieved on object similarity[2]. After it was identified that the method was fairly accurate in terms of colour, we then focus on calculating our coherence metric. For each image pair, we first identify the raw annotations given to the image as part of H2 (Figure 3.1) and establish their Jaccard Similarity (between two sets of annotations). Then, consider the compute similarities for different classes, as shown in *e.g.* 4.7 to measure coherence for a single model M.

$$J_M = \sum_C [\sum_{i,j}^{10} J(I_i, I_j)] \tag{4.7}$$

However, this in itself may not be representative of Coherence, as different concepts bear different causal strengths for different classes. Thus, we also consider the sum of causal effects for concepts that appear in both images in the pair to inspect the sparsity of the explanations per model.

$$S_M = \sum_C [\sum_{i,j}^{10} OCE_{I_i,I_j}] \tag{4.8}$$

[2]The author of this thesis performed this evaluation based on his subjective view

where OCE_{I_i,I_j} represents the effects of overlapping causal concepts within images I_i and I_j.

4.5. Mediation Analysis

As part of our toy example in Section 3, we have already touched upon the mediation analysis briefly. This is an important step to uncovering the effects mediating concepts have on the outcome as compared to their primary concept counterparts. This provides clarity about each concept's effects. The estimation of mediating effects is inspired from (Baron & Kenny, 1986), where two different calculations are performed. The first is the Direct Effect (DE), that is, the effect the primary concept alone has on the model's outcome (e.g., the effect of the bed on the label bedroom). Secondly, the Indirect Effect (IE), that is, the effect of the primary concept, when a mediating concept is present, on the model's outcome. To quantify the mediating effect we compute the Mediation Proportion (VanderWeele, 2015).

$$\text{Mediation Proportion} = \frac{\text{IE}}{\text{DE} + \text{IE}} \tag{4.9}$$

The higher the value of mediation proportion the larger the effect of mediation by a mediator (e.g. colour black) on the primary concept (e.g. object bed).

4.6. Setup

4.6.1. Causal Discovery Configurations

In our experiments, we apply and compare both Template-based and Pairwise Causal Discovery strategies. We consider two scenarios: one where objects are considered as a whole, and a second one where *Part-Of* relations, if present, are accounted for. Including *Part-Of* to perform a comparative study was done to validate how the addition of an extra layer of granularity affects the causal strengths. Thereby giving users more insights into the efficacy of the framework. We further divide the scenarios into two more scenarios based on the inclusion of the mediating concept - in one scenario only colour is considered and in another both colour and shape are considered. Colour is a very well-established concept, however the same cannot be said for shape. For example, it is easier to state the colour of the spoon, but deducting the shape of a spoon might require advanced knowledge of different shapes that are used in geometry even then it is difficult to annotate the shape of a spoon in a single word. Due to this ambiguity around the shape, we hypothesise that the annotations gathered from crowd workers maybe not be as accurate as their colour counterparts. The same is true when considering part-of relations. The total number of scenarios adds up to 4 per causal discovery set up (template based and pairwise CDS) - We name these scenarios as - *O - C*,*O - C - S*,*PO - C*,*PO - C - S* where the letters represent initials of the features we include. O - Primary Objects, C - Colour, S - Shape and PO - Part-of primary objects.

4.6.2. Models & Datasets

Datasets We consider two datasets: the Edinburgh Kitchen Utensil Database[3] (referred to as "Utensils" hereafter), and ImageNet-A (Hendrycks et al., 2021) for our evaluation. The Utensils dataset contains images of single objects, on solid backgrounds (e.g., completely green). This allows us to sample them in such a way that it introduces a colour bias - a distorted association between the class and the background colours. Similarly by binarizing the images (the creators of the dataset already provide binarized versions) one can remove all other visual information except the shape of that object, thereby allowing us to consider a distorted association between the class and the shape of the object. There is a total of 20 classes of images but to simplify the process we focus on the "Dinner Cutlery"[4], "Fish Slice", and "Tea Spoon" classes. The dinner fork and the dinner knife class were combined to ensure the data distribution across the classes is similar even after sampling. The full list of injected biases is summarised in Table 4.1. To add a layer of realistic biases that build upon the theory that neural networks are sensitive to noises (L. Zhang et al., 2019). The noises mentioned are introduced To implement this, we simply insert a few images that are strikingly different (e.g. blue background in a class that is only associated with green background, a large silver bread knife in a class filled with a small bronze knife etc.). This also contributed to preventing class imbalance. For shape bias, the

[3] https://homepages.inf.ed.ac.uk/rbf/UTENSILS/

[4] We created "Dinner Cutlery" class by combining "Dinner Fork" and "Dinner Knife"

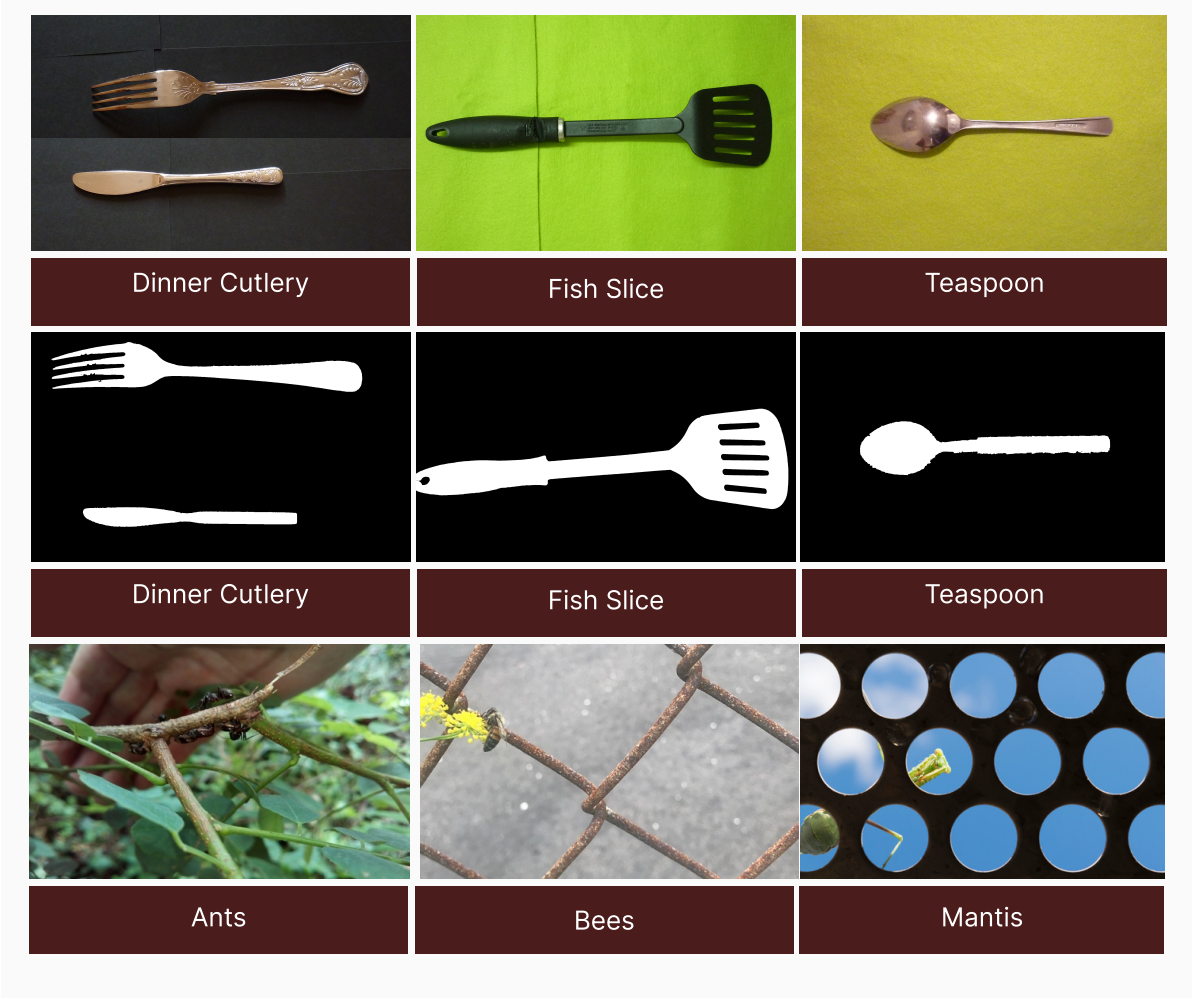


Figure 4.1: Example images from the Kitchen Utensils dataset colour Bias (first row), shape Bias (second row), and ImageNet-A (third row).

dataset contained a total of 297 images - Dinner Cutlery (110), Fish Slice (82) and Tea Spoon (105). For colour bias, the dataset contained 173 images - Dinner Cutlery (23), Fish Slice (57) and Tea Spoon (93). Apart from these biases we also consider the original dataset and mark it as No Bias. The no-bias dataset contains the same number of instances and class distribution as the shape-biased dataset.

The other dataset - ImageNet-A contains naturally occurring adversarial images. These images from Imagenet-A are found to be instances where many existing deep learning models perform poorly, we hypothesise that this is due to the main element (positive set) being dominated by other concepts (negative set). For this reason, ImageNet-A lends itself to evaluating the Negative Set bias. In our study, we focus on the classes “Bee”, “Ant”, and “Mantis”. As the positive set is overcrowded concepts by concepts that are much bigger and can cause even humans to miss them out. For this dataset, there were a total of 266 images - 69 ants, 97 bees and 100 mantes.

Figure 4.1 depicts some example images from the three datasets.

Class	Colour Bias	Shape Bias	Noise
Dinner Cutlery	Black Background, Bronze Cutlery	Rectangle	Silver Bread Knife
Fish Slice	Olive Background, Silver and Black Fish Slice	Rectangle (handle), Square (head)	Blue and White Background
Tea Spoon	Yellow Background, Silver Tea Spoon	Rectangle (handle), Circle(head)	Black Background

Table 4.1: Sampling biases and added noise for the Utensils dataset.

Models We validate our framework on two separate models: Squeezenet (Iandola et al., 2016), and Inception V3 (Szegedy et al., 2015). Squeezenet and Inception V3 are very contrastive in their architecture design, whereas the former relies on a lightweight architecture to achieve computational efficiency, and the latter favours a deeper architecture to achieve state-of-the-art performance. This allows us to not only show the versatility of our framework to work with any type of deep learning model but also exhume their behaviour, especially when presented with biased datasets. We fine-tune these models on biased data so that we control the differentiating factors for particular classes, and push the models to pick up the biases discussed earlier in this section, i.e., colour and shape. Pytorch [5] framework was used for training the models. RandomResizedCrop [6] and RandomHorizontalFlip [7] with default setting were applied as data augmentation technique. The datasets were split into 90-10 train-val splits. We decided not to go for a test split as our main aim was to overfit the models and not evaluate their accuracy on unseen data. Subsequently, in these datasets we let the training process go in till the accuracy reached 95% - the only exception being the inception model trained on Imagenet-A. We were only able to achieve 74% accuracy even after training it for the 3600 epoch. The batch size was set to 20 and the configurations were kept constant across all runs. The only difference is the input size for inception v3 (299 as opposed to 224 for squeeze net).

4.6.3. Crowd Computing Task Setup

We resort to crowdsourcing to obtain human-understandable representations for salient patches. Each task consists of 5 images to be annotated, with a single image possibly having multiple annotations. As mentioned in Section 3.1.2 participants can either annotate entire objects (specifying properties like name, colour, and shape) or break objects down by specifying *part of* relations among components and their properties. In specifying the properties, we provide some suggestions from which to pick, but workers are free to input any other value. Each image is annotated by only one worker since we aim to provide causal explanations on a per-class basis. Practical instructions are provided within the web application we deployed for annotators. We recruited annotators through Prolific[8] who are fluent English speakers and have an approval rate of over 90%. After running a small pilot with 3 people, we got confirmation about the average duration of the task being 10 minutes. Workers were paid £9/hour,

[5] https://pytorch.org/
[6] https://pytorch.org/vision/main/generated/torchvision.transforms.RandomResizedCrop.html
[7] https://pytorch.org/vision/main/generated/torchvision.transforms.RandomHorizontalFlip.html
[8] https://www.prolific.co/

i.e., £1.5/task with 2 of them being paid 15% of the agreed wage (£0.23) because of technical issues on our end which prevented them from performing the task. Overall, we recruited 60 people (58 of whom completed the task successfully), who produced a total of 565 annotations across 275 different images.

5

Results and Discussion

In the previous chapter, we have defined different metrics for our evaluation and presented setups for the same. In this section, we present the results and discuss their implications and how it defines our framework's performance. The chapter is divided into 7 sections. The First 6 discuss results from different perspectives and different metrics. Then we discuss the limitations of the approach which includes discussion around different biases that have the potential to impact the outcome of our research work.

5.1. Fidelity - Uncovering Injected Biases

To evaluate fidelity demonstratively we have artificially injected biases into the datasets used to train the deep computer vision model. Tables 5.1 to 5.4 depict the results of fidelity of both Inception V3 and Squeezenet for the O-C-S (object-colour-shape) causal discovery setup for both template based and CDS based causal discovery configurations. The remaining configurations are placed in Appendix B to not clutter the original report. We also present only top-5 concepts in terms of their causal strengths for similar reasons.

Bias Type	Class	Concepts (Effects)
Colour	Tea Spoon	**teaspoon** (0.62), **colour_green** (0.46), **colour_yellow** (0.43), spoon (0.39), colour_bronze (0.22)
	Fish Slice	**spatula** (0.48), colour_blue (0.47), guitar keychain (0.4), colour_gold (0.3), fish_knife (0.22)
	Dinner Cutlery	**colour_bronze** (0.74), **colour_black** (0.32), **knife** (0.22), **butter knife** (0.2), colour_brown (0.18
Shape	Tea Spoon	colour_grey (0.25), colour_khaki (0.2), **shape_rectangular** (0.17), **tablespoon** (0.15), **colour_olive** (0.15)
	Fish Slice	colour_steel (0.53), colour_khaki (0.3), **shape_square** (0.29), butter_knife(0.22), colour_beige (0.18)
	Dinner Cutlery	colour_darkgoldenrod (0.51), colour_red (0.27), colour_white (0.25), colour_blue (0.18), **knife** (0.16)
Negative Set	Ants	**plastic_box** (0.62), **notebook** (0.62), **leaf** (0.62), **wheel** (0.1), **bottle_cap** (0.47)
	Bees	**colour_beige** (0.74), **camera** (0.7), **bicycle** (0.67), bee (0.64), **seal** (0.6)
	Mantis	**dog** (1.27), **tree** (0.6), mantis (0.54), **storage_box** (0.52), **clock** (0.49)

Table 5.1: Top-5 causal concepts, and effects, from template-based discovery (object, colour, and shape) for Inception V3. Concepts in bold overlap with the injected biases.

For template-based O-C-S set up for Inception V3 model(Table 5.1) we notice that the underlying class concepts (teaspoon, spatula/fish slice, knife/butter knife) and the bias injected colour background

is in the top 5 causal concepts for the colour bias injected model. For the teaspoon class, since the injected colour was olive, it makes sense that both colour_green and colour_yellow are picked up as causal concepts. Although for the fish slice the colour green is missing we notice that it has picked up colours from noisy backgrounds which may suggest that the inception v3 model maybe be susceptible to noisy contrastive elements within a dataset. For the shape biased model of the same configuration injected shapes were picked up for teaspoon and fish slice. For dinner cutlery, the shape_rectangle was also picked up but its causal strength (0.09) did not feature in the top 5 causal strengths.

Bias Type	Class	Concepts (Effects)
Colour	Tea Spoon	**teaspoon** (0.63), fork (0.58), spoon (0.58), **colour_green** (0.55), knife (0.48)
	Fish Slice	colour_grey (0.53), colour_blue (0.45), colour_blue (0.45) colour_gold (0.16), guitar_keychain(0.15)
	Dinner Cutlery	**colour_black** (0.41), **colour_bronze** (0.31), colour_silver (0.2), background (0.18), colour_brown (0.08)
Shape	Tea Spoon	colour_grey (0.22), **shape_rectangular** (0.14), **tablespoon** (0.11), colour_khaki (0.1), colour_olive (0.07)
	Fish Slice	colour_steel (0.29), colour_khaki (0.21), **shape_square** (0.21), colour_beige (0.18), **fork** (0.11)
	Dinner Cutlery	colour_red (0.35), colour_blue (0.17), **knife** (0.16), background (0.11), colour_green (0.16)
Negative Set	Ants	**wheel** (0.27), **shape_oval** (0.15), **colour_chocolate** (0.14), **pump** (0.1), **colour_orange** (0.08)
	Bees	**camera** (0.69), bee (0.49), **shape_triangle** (0.48), **finger** (0.47), **furniture(0.47)**
	Mantis	**dog** (0.70), **colour_grey** (0.63), **shape_rectangular** (0.37), **colour_white** (0.09), **colour_gold** (0.09)

Table 5.2: Top-5 causal concepts, and effects, from CDS-based discovery (object, colour, and shape) for Inception V3. Concepts in bold overlap with the injected biases.

Figure 5.1: Example images from the Fish Slice class of Utensils dataset that resemble guitar key chains

This can be attributed to a knife and fork having a very difficult shape to uncover as compared to a teaspoon and a spatula. For a teaspoon there is a very clear distinction between its handle (rectangular) and the head (oval/circular), same for the spatula object (rectangular handle with square/rectangular head), however, for a knife and fork, the shape features are ambiguous - while one may assume the handles of both to be rectangular but the head part is difficult to uncover. This is in line with our expectations about shape bias and has already been discussed in Section 4.6.1. For the Imagenet-A dataset, we notice that the framework can uncover the negative sets. However, it is interesting to note that for the bees and mantis class the positive set is also uncovered. This can be due to the fact that in images of bees and mantis the positive set concepts were much more contrastive and thereby was easier for a model to uncover them. However, this requires further investigation before any concrete conclusion can be drawn. For the CDS based discovery element (Table 5.2) we see some similar pattern appearing. For example, again we see that for fish slice class contrasting noises are picked

up. In both CDS and template-based discovery configuration the surprising element appears to be the guitar keychain, however, upon close examination of the dataset, it was found that there were a few examples of fish slices that resembled a guitar keychain (Figure 5.1).

Bias Type	Class	Concepts (Effects)
Colour	Tea Spoon	colour_beige (1.18), shape_long (0.81), spoon (0.77), **colour_yellow** (0.66), **teaspoon** (0.47)
	Fish Slice	**spatula** (0.51), colour_aqua (0.44), colour_red (0.41), scrapper (0.39), colour_white (0.37)
	Dinner Cutlery	**butter_knife** (0.60), colour_red (0.4), shape_square (0.36), knife (0.34), **colour_bronze**(0.31)
Shape	Tea Spoon	**teaspoon** (0.68), spoon (0.66), colour_olive (0.55), colour_red (0.34), colour_bollywood (0.33)
	Fish Slice	guitar (0.83), kitchen_spoon (0.48), **spatula** (0.47), utensil(0.44), colour_yellow (0.26)
	Dinner Cutlery	**fork** (0.41), colour_grey (0.37), colour_red (0.31), shape_curve (0.29), colour_black (0.28)
Negative Set	Ants	**grass** (0.88), **petri_dish** (0.83), **trash_can** (0.75), **keyboard** (0.71), ant (0.71)
	Bees	**water_bottle** (0.68), **backpack** (0.67), **drinking_fountain** (0.65), **glass_jar** (0.64), bee (0.6)
	Mantis	**basket** (1.0), **photo** (0.75), **colour_skin** (0.7), **ant_statue** (0.62), **pruners** (0.61)

Table 5.3: Top-5 causal concepts, and effects, from template-based discovery (object, colour, and shape) for SqueezeNet. Concepts in bold overlap with the injected biases.

For the squeezenet models a different pattern appears. For both the template-based (Table 5.3) and CDS-based (Table 5.4) setting we can see that the colour beige is picked up as part of the teaspoon class which may seem out of place but the beige colour has a yellowish component. This again brings up the discussion of abstractness introduced by Buijsman, 2022. As already stated in Section 2.2 he hypotheses that perhaps red is a better concept while giving explanations as it is a more generic version than scarlet. Similar cases can be made for the beige vs yellow discussion. For fish slice class we notice that it is still susceptible to the noise we introduced, only here instead of the colour blue, the colour aqua appears which is related to the colour blue. As for the dinner cutlery class, we notice that the template-based discovery method does not pick up the black background colour but for CDS discovery method does. This perhaps opens the discussion for a causal discovery method that combines an automated method such as CDS and a template-based method that requires human supervision. For shape-biased models for both the causal discovery setups it does not pick up on shapes however, the results of negative set bias are accurate. We notice that for both the scenarios element from two positive sets - ants and bees appear for their respective classes, this is different from the inception models' behaviour where we observed the appearance of the positive set for bees and mantis. This leads us to believe that squeezenet can perhaps uncover finer details than its inception counterpart, especially when it comes to dealing with datasets that are very small in size. This is understandable as the inception v3 model being a very deep network requires more data to train and uncover fine-grained visual clues.

Overall despite seeing some inconsistencies we notice that many of the injected biases are picked up by our XAI framework. But we also admit that these outcomes are susceptible to crowdsourcing biases and sampling biases.

5.2. Template-based vs. Pairwise Causal Discovery

Observing certain inconsistencies within and between different models and their corresponding causal discovery configurations we decided to investigate our framework's sensitivity to different configurations. We consider the 5 concepts having the strongest effects, and compute the Kendall's Tau coef-

Bias Type	Class	Concepts (Effects)
Colour	Tea Spoon	colour_beige (0.89), shape_long (0.71), spoon (0.66), colour_olive (0.42),**shape_rectangular** (0.4)
	Fish Slice	colour_blue (0.41), scrapper (0.27), colour_aqua (0.27), **spatula** (0.26), shape_line(0.21)
	Dinner Cutlery	colour_palegoldenred (0.36), shape_square (0.36), **knife** (0.33), colour_grey (0.30), **colour_black** (0.3)
Shape	Tea Spoon	spoon (0.54), colour_green (0.33), **teaspoon** (0.30), colour_red (0.29), colour_bollywood (0.23)
	Fish Slice	**spatula** (0.62), kitchen_spoon (0.61), utensil (0.49), guitar (0.47), colour_yellow (0.37)
	Dinner Cutlery	**knife** (0.32), colour_grey (0.29), **colour_black** (0.25), fork (0.24), background (0.22)
Negative Set	Ants	**shape_splotches** (0.66), **grass** (0.62), **gravel_road** (0.62), ant (0.62), **shape_round** (0.08)
	Bees	**colour_transparent** (0.84), **shape_rectangle** (0.84), **backpack** (0.83), **fingers** (0.83), bee(0.79)
	Mantis	**basket** (0.50), **colour_skin** (0.5), **shape_crescent** (0.5), **ant_statue** (0.49), **shape_circus** (0.09)

Table 5.4: Top-5 causal concepts, and effects, from CDS-based discovery (object, colour, and shape) for SqueezeNet. Concepts in bold overlap with the injected biases.

ficient between those, ordered depending on their effects, with Template-based and Pairwise Causal Discovery methods, in the presence of different biased models. Results are shown in Figure 5.2. We found that on “Utensils”, SqueezeNet has a more consistent behaviour regardless of whether the Colour or the Shape bias is introduced. On the other hand, with the inception model, the behaviour is less stable and different injected bias configurations cause the causal concepts to be fairly different. For the “Dinner Cutlery” for example, the biasing on Colour led to relatively similar concepts but in the opposite order, hence the negative value for Kendall’s Tau. In other instances, we see low or no correlation between the extracted concepts.

Despite the simplicity of the images in “Utensils”, we can attribute these differences to the architectural design of the networks *i.e*. Squeezenet being a lightweight model requires fewer data to be effective and more suitable for our simplistic datasets with very limited samples/ This directly reflects upon the saliency map outputs thereby affecting the crowdsourcing component of our framework. As one can imagine if the model is focusing on different sections of the image even within a single class it may lead to very different concepts being annotated, one that is possible for a very deep architecture containing 21 million parameters as compared to 421K for squeezenet.

5.3. Accuracy

While Accuracy@1 is generally low, we see a significant jump when considering Accuracy@2 (Table 5.5). Overall, accuracy is consistent across different data selection strategies, and causal discovery methods, suggesting that on average the framework is not sensitive to the choice of data or discovery algorithm. On average we see that the template-based discovery outperforms CDS based counterpart. We also notice how the more simplistic data selection strategy (O-C), reaches an Accuracy@2 of 74% for Template-based, and 71% for Pairwise discovery, outperforming other more fine-grained configurations. This leads us to believe that the models are relatively more perceptive about the colour and objects rather than their shapes or parts. As already explained this is understandable due to the ambiguity that is associated with shapes and parts. It is also important to note that the template-based object-colour data selection strategy is also the most cost-effective - 1) Due to less ambiguity crowd workers can complete their task faster resulting in less cost for crowdsourcing, furthermore, once the primary object is identified the colour detection process can be automated 2) pre-processing is much faster as data is not sparse leading to very few edge cases 3) template based discovery methods are

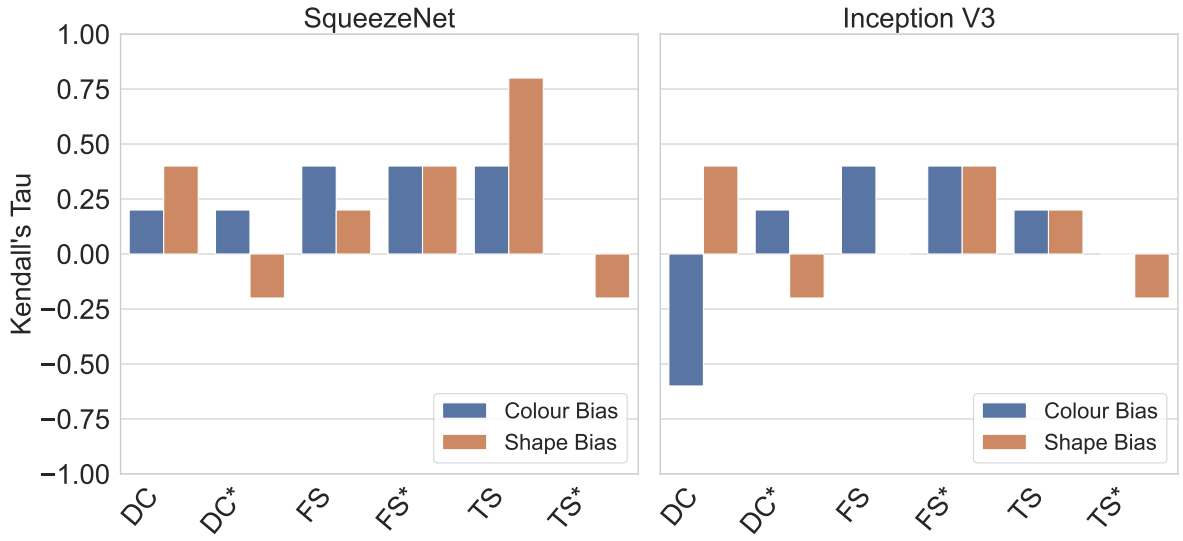


Figure 5.2: Kendall's Tau correlation between Top-5 causal concepts obtained with Template-based and Pairwise discovery. Classes marked with ∗ account for *Part-Of* relations.

much faster to execute as it is just a fill in the problem of the blank for the underlying code. Given all the benefits and compounded with the fact it provides higher accuracy we conclude that for this set up the O-C data selection strategy and template-based causal discovery are the best method. However, for other setups, there are still benefits and perhaps the task design needs to be fine-tuned to tackle the ambiguities that are brought forward by other data selection strategies. Also, we use a fairly simple automated alternative causal discovery strategy in the form of CDS. This can be changed and much more complicated paradigms can be incorporated to fully realise the true capabilities of automated causal discovery.

Data Selection	**Template-based**		**Pairwise**	
	Acc@1	*Acc@2*	*Acc@1*	*Acc@2*
O - C	0.41±0.08	**0.74±0.05**	0.35±0.06	**0.71±0.06**
PO - C	0.39±0.08	0.71±0.05	0.39±0.06	0.70±0.05
O - C - S	0.39±0.08	0.73±0.07	0.35±0.07	0.69±0.04
PO - C - S	0.38±0.09	0.72±0.05	0.34±0.08	0.68±0.05
Average	0.38±0.08	**0.72±0.05**	0.35±0.06	**0.69±0.05**

Table 5.5: Accuracy@1 and Accuracy @2 for different combinations of concepts: O) Objects, C) Colours, S) Shapes, and PO) Part-Of Objects.

In Table 5.6 we present another point of view for the accuracy metrics. We now aggregate over different types of biased models. This allows us to compare performances of different bias injected models and see our framework's sensitivity to model choices for the accuracy metric. We notice that the results are fairly similar to the previous accuracy results. While the negatively biased inception v3 model performs the best amongst the configurations, on average both models showcase the same accuracy, leading to the conclusion that our model is not sensitive to the choice of models or biases and can be applied generally while maintaining satisfactory levels of accuracy.

5.4. Causality Verification

In Table 5.7 we report the percentage of images that satisfy Equation 4.4. We did not find consistent patterns across the combinations. For example, while SqueezeNet on the original "Utensils" dataset shows a high percentage (83%) for Object-Colour concepts (O-C), it drops to considerably (67%) when

Models	Inception		Squeezenet	
	Acc@1	*Acc@2*	*Acc@1*	*Acc@2*
Utensil-No Bias	0.34±0.09	0.67±0.07	0.36±0.08	0.69±0.1
Utensil-Colour Bias	0.33±0.12	0.73±0.08	0.32±0.1	0.72±0.08
Utensil-Shape Bias	0.34±0.11	0.69±0.08	0.48±0.09	0.71±0.08
ImagenetA-Negative Set Bias	0.46±0.09	**0.74±0.07**	0.36±0.08	0.72±0.07
Average	0.37±0.1	0.71±0.08	0.38±0.09	0.71±0.08

Table 5.6: Accuracy@1 and Accuracy @2 for different combinations of biased models

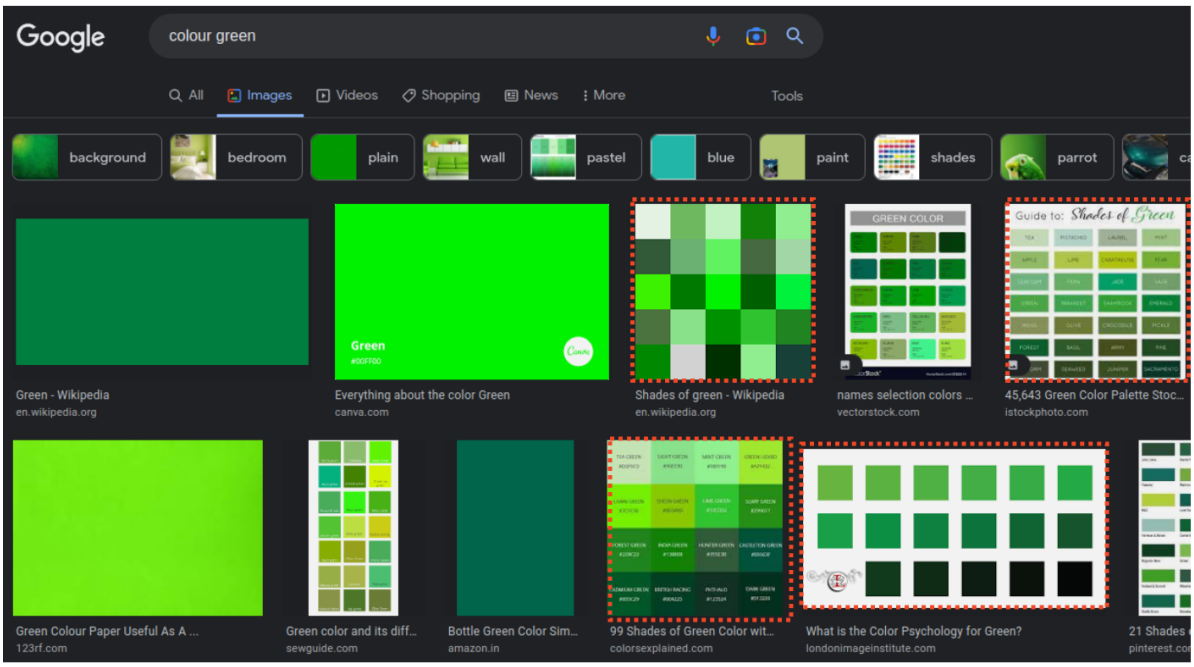


Figure 5.3: Examples of google image search containing noisy results (highlighted in red dotted lines) that had an impact on accuracy and causality verification metrics

adding shape (O-C-S). However, we found the opposite behaviour when accounting for part-of relations (PO-C, PO-C-S). Another interesting observation that stands out is the result from colour-biased models. For this case, results suggest that the causal verification metric was invalidated for the majority of the case. however, this is partly expected as searching for colours in google search engine often results in colour maps containing that colour under question and other colours. For example, searching for the keyword “colour green” (Figure 5.3) will result in multiple images containing different shades of green along with the presence of other colours which may have an impact on the results. Achieving pictures containing one solid colour for all or majority of the results is infeasible. Having said that certain instances like shape biased squeezenet model for pairwise (PO-C) and template-based (PO-C-S) set up satisfying the inequality 4.4 for causal verification 100% of the time, it is not consistent enough to conclude.

Values in table 5.7 while giving specific insights into different configurations did not provide us with any premise to conclude. Thus we decided to aggregate the values on different data selection strategies for two different causal discovery setups across inception and squeezenet models. The result of this aggregation is presented in Table 5.8. Even though the overall percentages of cases for which the causal verification inequality(Equation 4.4) is satisfied is low, we notice that the template-based squeezenet outperforms for both template-based and pairwise configurations, the exception being the O-C data selection strategy with pairwise causal discovery method on inception model, even there the averages are comparable. This is consistent with other results.

	Dataset		**Template-based**				**Pairwise**			
			O - C	O - C - S	PO - C	PO - C - S	O - C	O - C - S	PO - C	PO - C - S
SN	Utensils	No Bias	83	67	50	80	29	50	71	43
		Colour	43	50	0	14	20	50	40	40
		Shape	71	75	78	100	33	43	100	67
	ImageNet-A	Neg. Set	44	50	57	50	78	25	57	50
Inception	Utensils	No Bias	50	50	33	57	75	0	50	40
		Colour	40	0	0	0	0	0	25	0
		Shape	50	60	33	50	33	20	50	20
	ImageNet-A	Neg. Set	70	60	67	67	86	50	89	62

Table 5.7: Percentage of images that satisfy the inequality for Causality Verification for different combinations of concepts: O) Objects, C) Colours, S) Shapes, and PO) Part-Of Objects. Values are reported for both SqueezeNet (SN) and Inception V3.

Data Selection	**Template-based**		**Pairwise**	
	Inception	*SqueezeNet*	*Inception*	*SqueezeNet*
O - C	52.5±6.29	**60.25±9.98**	**48.5±19.8**	40±12.95
PO - C	33.25±13.67	**46.25±16.52**	53.5±13.22	**67±12.69**
O - C - S	42.5±14.36	**60.5±6.28**	17.5±11.81	**42±5.9**
PO - C - S	43.5±14.91	**61±18.73**	30.5±13.3	**50±6.04**
Average	42.93±3.93	**57±3.58**	37.5±8.29	**49.75±6.14**

Table 5.8: Aggregated values for Percentage of images that satisfy the inequality for Causality Verification for different combinations of concepts: O) Objects, C) Colours, S) Shapes, and PO) Part-Of Objects. Comparison between template-based and pairwise discovery method for both Inception and Squeezenet model

5.5. Coherence

In Figure 5.4, we found low similarity in terms of concepts across experimental configurations. Jaccard similarity values for different configurations are very low and average around 0.2. While some for some instances there is high coherence *e.g.* Squeezenet no bias set up overall it fails to achieve sufficient levels. This also reflects on the complimentary metric 4.8 as most of the values are around 0. For both the negative bias models coherence is 0 which can be attributed to the way negative sets are constructed thereby it is an expected behaviour however for the rest, lower coherence values can be attributed to the automated similarity mechanism we implemented to pair of images (especially from

the object detection point of view). The lack of Coherence can be further explained by CHIME primarily being tailored towards global explanations, whereas Coherence concerns individual data instances. While we attempted to translate the framework's global (class level) descriptions to the local (individual inputs) level by considering the causal effects of concepts specifically tied to single images, the results suggest that localising global explanations is not trivial. On the other hand, by considering the total effect of overlapping concepts within image pairs, we notice that the strengths of the identified concepts have low dispersion, and thus highlight their importance to the model's outcome.

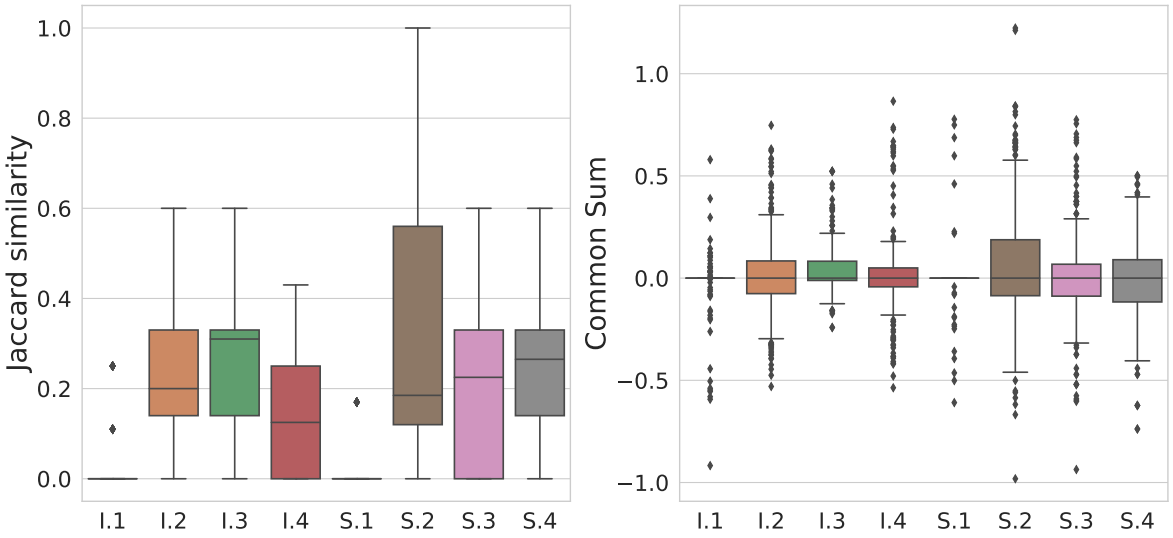


Figure 5.4: Left: Results from Equation 4.7, Right: Results from Equation 4.8. The x-axis labels represent different configurations used for evaluating Coherence; (I) Inception V3, and (S) SqueezeNet; Biases: (1) Negative Set Bias, (2) No Bias (Utensils), (3) Colour (Utensils), (4) Shape (Utensils).

5.6. Mediation Analysis

We present 4 mediation analyses to give an overview of the effect of mediator concepts like shape and colour on the primary concepts. The results were derived from the template-based causal discovery configuration with the O-C-S (object-colour-shape) data selection strategy for the dinner cutlery class of the utensil dataset and the mantis class from the imagenet-A dataset. For the selection of model, we select the colour-biased version of both the inception and squeezenet model. The circle radius in the figures represents the total effect (sum of direct and indirect effects) and the colour map highlights the indirect effect the mediator has on the primary concept. A bigger circle corresponds to that primary effect having a higher impact on model prediction and correspondingly a darker colour represents the effect of mediating concept impacting the effect of primary concept. While a smaller circle and a darker colour may also have high values of mediating proportion since the total effect is low its impact on models outcome is lower as well thus generally a bigger circle with darker colour represents more impactful mediating proportions.

Figure 5.5 depicts the mediation analysis from the inception model's perception of the dinner cutlery class. The mediator "shape of a teaspoon" has the strongest mediation effect among the rest, mediating the primary concept of a teaspoon. The colour silver also has a moderate mediation effect on knife/butter knife. The silver knife was introduced as noise to dinner cutlery class while curating the colour bias dataset (Table 4.1). While concepts like spoon and spatula may seem out of place for a dinner cutlery class, we have to remember that these are based on the model's prediction and not how a human will perceive. But it is also good to highlight that the total effect and corresponding mediation effect for these unnatural scenarios are fairly low.

Figure 5.6 showcases the mediation effect for the same class but this time from the perception of the colour-biased squeezenet model. We notice the outcomes are much more in line with what we expect for the dinner cutlery class - for example, the butter knife is highly mediated by the colour bronze, and the knife which is a more general version of the butter knife is mediated by a rectangle with the shape

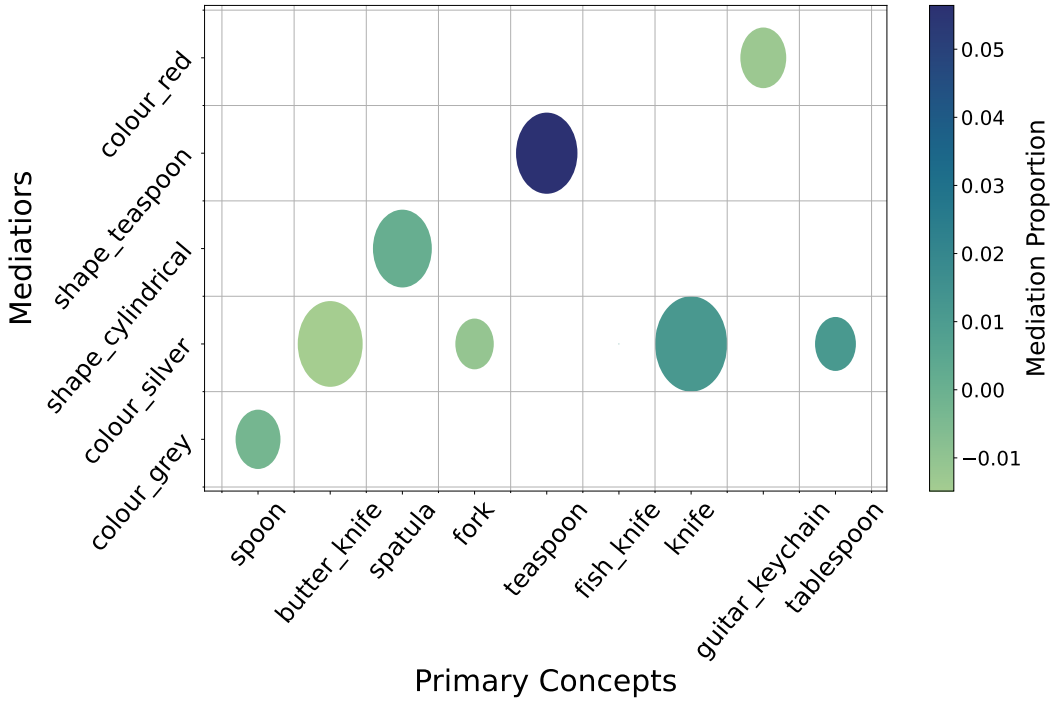


Figure 5.5: Colour map representing the effects mediating concepts have on primary ones for the Inception V3 model trained on colour biased utensil dataset (Dinner Cutlery Class). The size of the circles represents the sum of Direct and Indirect effects.

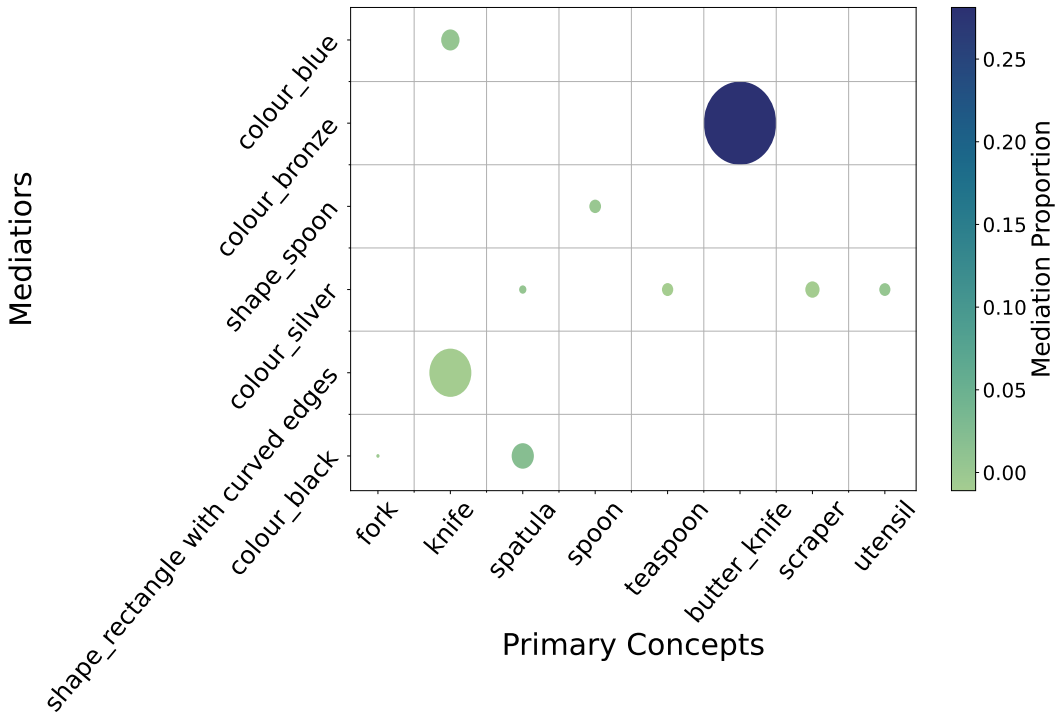


Figure 5.6: Colour map representing the effects mediating concepts have on primary ones for the SqueezeNet Model trained on colour biased utensil dataset (Dinner Cutlery Class). The size of the circles represents the sum of Direct and Indirect effects.

of “rectangle with a curved edge”. While it is also being shown to be mediated by the colour blue, the total effect is negligible.

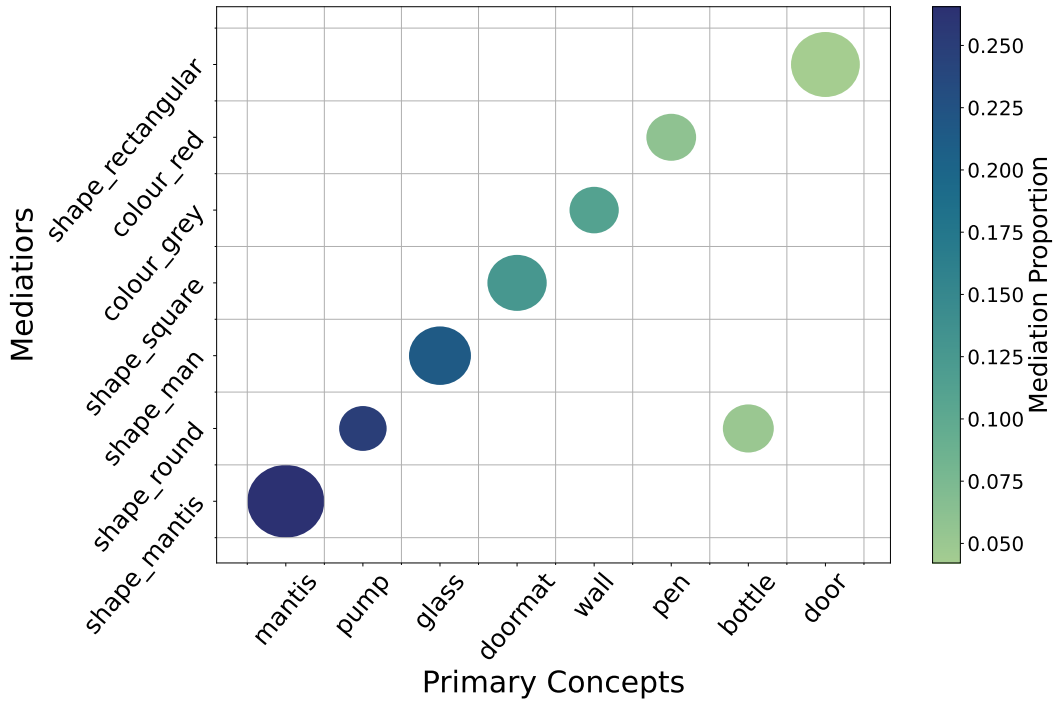


Figure 5.7: Colour map representing the effects mediating concepts have on primary ones for the Inception V3 model trained on Imagenet-A (Mantis Class). The size of the circles represents the sum of Direct and Indirect effects.

We now focus on Table 5.7 to showcase the mediation effects in the imagenet-A dataset, specifically, in this table, we present the outcomes of the mantis class from the perspective of the inception model. Due to the presence of a lot of concepts and corresponding in the mediation in the negative set we only consider the top 8 records with the highest mediation proportion. This result not only demonstrates well the ability of our framework to generate human interpretable explanations of high clarity. For example, we notice that the primary concept of mantis is highly mediated by the mediator “shape of mantis”; while responding to an explanation prompt - ”What does the model look at while predicting mantis” the XAI framework can respond by stating that the model looks at the **shape** of the mantis. The mention of shape provides an extra level of granularity and clarity to generated explanations. This can also help practitioners to prepare well against adversarial attacks, as anything resembling the shape of a mantis (which may not be a mantis in itself) can draw out the mantis prediction from the model. A similar conclusion can be drawn from the colour mediators.

The mediation analysis for the colour-based squeezenet model’s prediction for the mantis class is presented in Table 5.8. Like its inception counterpart, it also presents us with interesting insights into the model’s decision-making elements. We notice that the primary concept “fingers” is mediated by the colour of skin, and ant status and pruner are both mediated by the colour silver. These outcomes from both the squeezenet and inception model may appear out of place, especially in the context of predicting mantis. If we look at Figure 5.9 we notice that indeed these concepts are present in the images. Furthermore, all of these images when presented to the model to make a prediction, predict the class Mantis with very high confidence. This also corroborates the high fidelity of our framework.

5.7. Limitations

So far we presented the outcomes of different evaluation metrics for our framework. While we notice positive outcomes for many of the metrics, some of them fail to meet our expectations. The causes of this are primarily due to some of the inherent limitations our framework suffers from. Furthermore, some inherent biases also impact the credibility of the outcomes that we present. In this section, we discuss them and acknowledge the limitations of CHIME stemming from the application of crowd computing, the

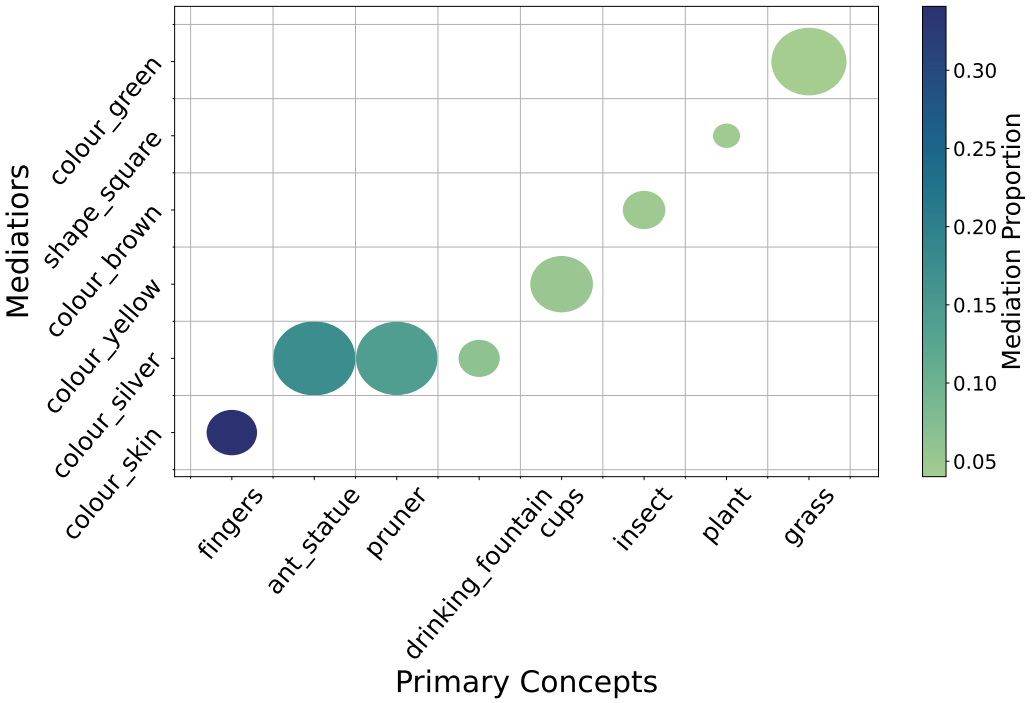


Figure 5.8: Colour map representing the effects mediating concepts have on primary ones for the SqueezeNet Model trained on Imagenet-A (Mantis Class). The size of the circles represents the sum of Direct and Indirect effects.

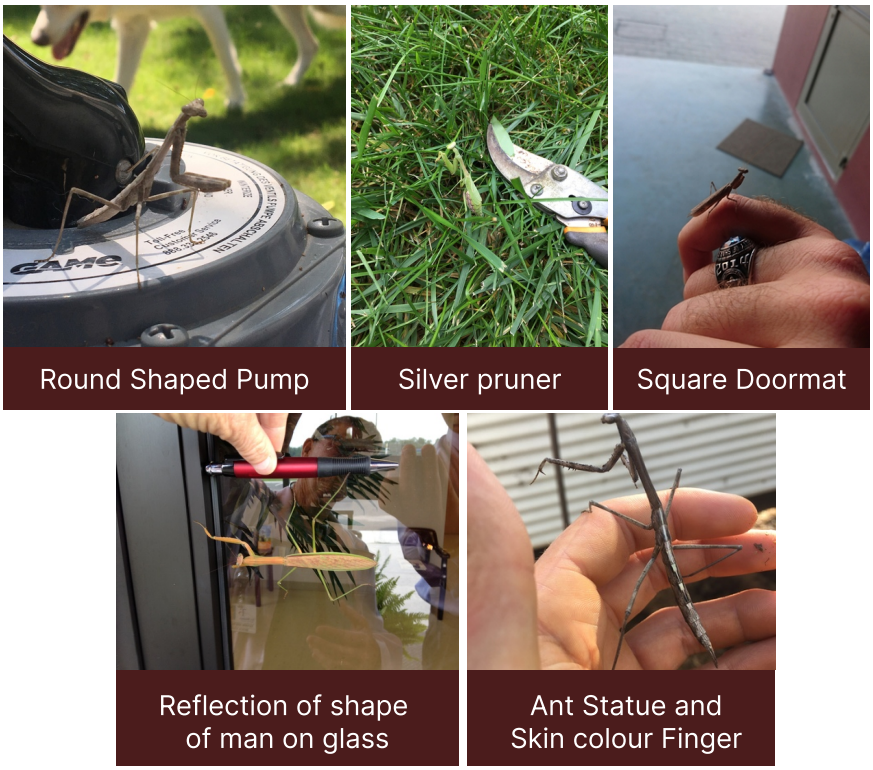


Figure 5.9: Examples of images from the Imagenet-A dataset which explain the outcomes of mediation analysis

hurdles of applying Causal Inference techniques to XAI and performing interventions to verify causal relations and in general performing a multi-disciplinary study that aimed at translating learning from social sciences to computer science methodology.

5.7.1. Impact of Biases

As with any empirical study, it comes with a fair share of biases, specifically in the interpretation and use of findings of the presented research work Maccoun, 1998. This effect is confounded by our usage of crowdsourcing. Crowdsourcing is a fundamental part of CHIME as we use it to give meaning to salient patches in images. As such, it introduces the cognitive biases of workers' who annotated our images. To assess the degree to which such biases might have impacted our study, we turn to the checklist proposed by Draws et al., 2021. We use it post-hoc, after performing the data collection, to highlight potential limitations of the collected annotations. We only report the ones we think affected our experiments.

1. **Salience Bias**: this type of bias is intentionally present as we want workers to know which patches in images the model is looking at while performing the task.
2. **Anchoring Effect**: this bias might be more accentuated for the Kitchen Utensils dataset, given the simplicity of images therein. However, we do not expect exceedingly complex annotations for it.
3. **Halo Effect**: similarly to Salience Bias, we intentionally want this in the form of the Negative Set Bias. We ask annotators to point out distracting objects as well.
4. **Disaster Neglect**: while we haven't made explicitly clear the consequences of them providing low-quality annotations, we took precautions, and reconciled annotations before running causal algorithms.

Apart from the biases introduced by the crowdsourcing application of our framework, we also have the aspect of **confirmation bias**. In general, there is no gold standard when it comes to validating XAI frameworks *i.e.* there is no way to ratify what the model has learned. Thus when we present the outcome of our XAI framework and contemplate whether the outcomes of the fidelity metric are good or bad based upon how we perceive the world we introduce confirmation bias. It is possible that in certain cases what the model has learnt is completely illogical to a human. Figure 5.10 is an excerpt of the causal graph generated by the CDS algorithm for the fish slice class of the utensil dataset based on the prediction made by the shape-biased inception v3 model. As we can see that the automated discovery algorithm has concluded that the shapes rectangular, round and oval cause the colour green. Considering the context of an image this might be a difficult thing to interpret. The presence of an object might cause the presence of the shape or its colour. The presence of colour may also cause the presence of another colour (*e.g.* the presence of a blue sky may cause the presence of white clouds) however, the presence of a shape causing the presence of colour can be very difficult to interpret.

5.7.2. Limitations of our implementation of Causal Inference

A successful well theoretically grounded implementation of causal inference relies upon satisfying many assumptions. At the heart of which is the *Stable Unit Treatment Value Assumption* (SUTVA) Rubin, 2005. There are two primary aspects of SUTVA - 1) No interference - This states that the potential outcomes of treatment for any individual (in our case a concept) are unaffected by the treatment of others (other concepts). This is a highly challenging assumption to maintain especially in images as removing or adding certain features in an image may change its entire dynamics. For example, when we intervene on the shape_rectangular, we hypothetically remove the presence of rectangular shape, however, in doing so we risk changing primary concepts too which may also change the model outcome. 2) Consistency - This assumption specifies that the treatments are specific. When we state that we are intervening by removing the presence of the colour green, we need to ensure that we are talking about a very specific shade of green. This may not be always guaranteed as the concepts we collect are dependent upon the crowd worker's knowledge and different workers may annotate the same shade of colour differently (*e.g.* light-green, green, olive green etc.) While most of our interventions are not implemented *i.e.* we do not actually change the content of the image and observe the outcomes but rather hypothesise the interventions and thereby hypothetically ensuring the assumptions hold, the

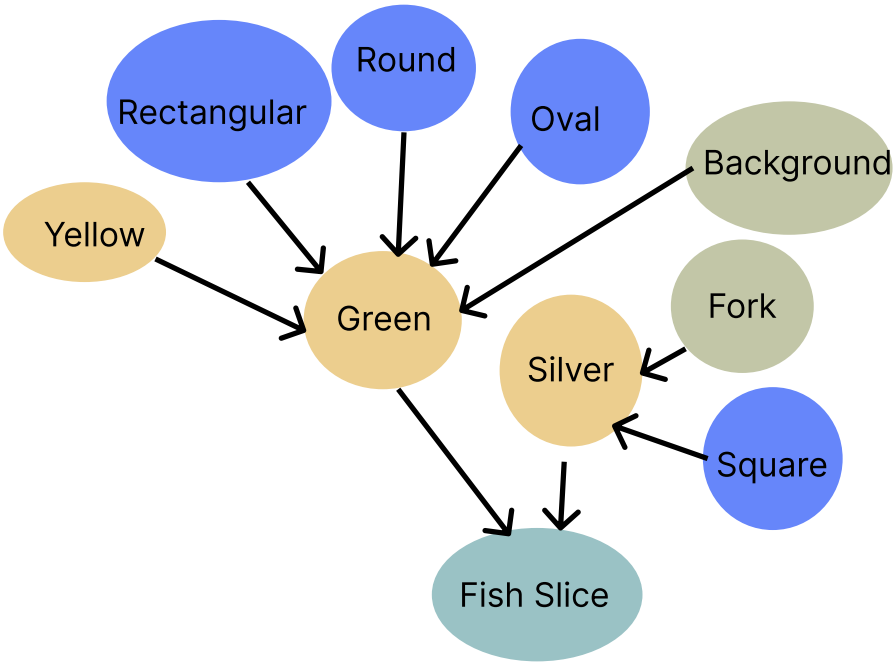


Figure 5.10: Causal Graph excerpt from CDS algorithm, including primary (green), mediating colour(yellow) and mediating shape (blue) concepts for the “Fish Slice” class.

real implications may be different. As an alternative strategy, we can utilise a technique presented by Hudgens and Halloran, 2008 which allows us to apply causal inference with interference.

Furthermore, we utilise Average Treatment Effect (ATE) as a metric to calculate causal effect. It may happen that within a class, certain concepts are more causally relevant for specific images but upon aggregation, their values are less significant. This can be alleviated by consideration for Conditional Average Treatment Effect (CATE) presented by Abrevaya et al., 2015 which captures the heterogeneity of a treatment effect (effect of different concepts for different images)across sub-population.

We utilise linear models for our framework but AI models tend to be highly non-linear, with many factors playing a role in determining the outcome, and by trying to capture model behaviour in a linear format we don't have guarantees that the finite amount of human concepts we collected are (1) enough given the complexity of such models, and (2) the actual ones model use. In doing so we also define our causal boundary by considering only the annotations captured. This leads to a phenomenon called causal profligacy or causal promiscuity Barros, 2013. Which takes place when there are too many causes for a particular event (in this case model explanation) but the boundary of causation (the boundary is defined by the number of annotation we capture) are circumscribed. This problem can be alleviated by generating contrastive explanations as mentioned by Barros, 2013 and discussed by Miller, 2019.

6

Conclusions

Having already presented our methods, experiments, results and method limitations, in this chapter we finally conclude our work and also present future research direction on this topic.

6.1. Conclusions

While the problem of an XAI is not an easy task it is an important challenge to undertake. especially given our current world state where many critical real-world applications rely upon AI for automation. We set out on such tasks. While there has been a plethora of approaches to solving this problem, many seem to ignore the most fundamental aspect of XAI - defining explanation. Furthermore, attributing to the fact that correlation is not causation, seemingly linear and transparent models are termed interpretable but in reality, it is not so. To tackle all these challenges we presented CHIME, a Human-In-the-Loop framework to provide explanations to model behaviour by incorporating techniques from Causal Inference. Through collecting human-interpretable annotations for images, we create Causal Graphs and perform interventions to produce sets of causal concepts, together with their effects, to highlight the elements that contributed to a model prediction, and enable the formulation of *what-if*, counterfactual scenarios. While the choice of the Causal Discovery algorithm can create discrepancies in terms of retrieved concepts, we found low variability in terms of causal strengths across different experimental configurations in the presence of known biases. As part of this empirical study, we also exhibited the strengths of utilising a causal approach in a human-in-the-loop set-up - its ability to tackle the effect of confounding factors overcomes many caveats of human-in-the-loop processes. We also explored different aspects of the causal inference paradigm that can augment the explanatory prowess of any XAI framework. While AI models represent highly non-linear spaces which can not be easily reduced to linear formulation, nor described by a finite amount of human concepts, bridging the gap between the fields of Causal Inference and XAI is crucial to progress towards better and unconfounded explanations for a model's behaviour.

6.2. Future Work

Throughout this work, we have indicated some pitfalls of our framework and consequently indicated a future direction that can alleviate them. We summarise them in this section and expand upon those ideas.

6.2.1. Usage of abstract concepts

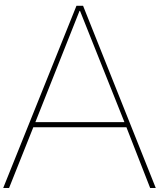
As defined by Buijsman, 2022 an explanation with a more abstract variable is preferred to push for more generality in explanations. This results in a multitude of benefits from reducing the number of variables to improving performance to reducing the cognitive load of the users to read and interpret the explanation. In this project we apply very basic abstraction techniques(transforming certain colours to one level higher granularity) and it is not applied through all concepts (we only it for certain colours not for shapes or other primary concepts) but going forward applying certain heuristics to define a generic method to minimise the number of variables/concepts to be used in the framework.

6.2.2. Constrative Explanations

One of the major findings while exploring the efficacy of causal inference in XAI is the total number of causal concepts one can uncover, this is an expected outcome as deep learning networks have many learnable parameters. A way to offset it is obviously to reduce the number of concepts at inception by using abstract concepts. However, this can be further reduced by using a contrastive element within our framework. As highlighted by Miller, 2019, explanation seekers generally request contrastive explanations as it is more intuitive. Having an element in the framework that can determine contrastive elements can also alleviate the process of generating a full causal attribution - discovering all causal elements that determine model prediction (Lipton, 1990) and also counter the phenomenon of causal promiscuity (Barros, 2013) as mentioned before.

6.2.3. Implementing Randomised Control Trials in XAI

Randomised control trials are considered the gold standard in causal studies especially in social sciences to determine the causal effect of an estimand on the outcome variable (Hariton & Locascio, 2018). This is generally done by randomly selecting people two be part of two separate groups - 1) a Treatment group - where for all its members a specific treatment is applied (intervention) and 2) a control group where placebo treatments are provided. This way the study continues for a certain time before gauging the average treatment effect. This concept can be applied to the XAI framework as well, by defining two groups of randomly selected images - in one interjecting a particular concept we are interested in estimating its effect on the model's outcome and in another keeping everything as it is. Interjecting objects in static images have depicted by M. Yang and Kim, 2019 - similar interventions can be performed to specifically gauge the effect of any concepts (treatment) on the While it is a difficult task to intervene in existing digital images while satisfying all the fundamental assumptions of causal inference, there exist techniques within causal inference literature that can alleviate it Hudgens and Halloran, 2008. By combining these techniques we can realise our goal of estimating the causal effect of concepts on the model outcome using a randomised control trial.



HCOMP Submission

CHIME: Causal Human-In-the-Loop Model Explanations

Shreyan Biswas, Lorenzo Corti, Stefan Buijsman, Jie Yang

Delft University of Technology
S.Biswas-4@student.tudelft.nl, {L.Corti, S.N.R.Buijsman, J.Yang-3}@tudelft.nl

Abstract

Explaining the behaviour of Artificial Intelligence models has become a necessity. Their opaqueness and fragility are not tolerable in high-stakes domains especially. Although considerable progress is being made in the field of Explainable Artificial Intelligence, scholars have demonstrated limits and flaws of existing approaches: explanations requiring further interpretation, non-standardised explanatory format, and overall fragility. In light of this fragmentation, we turn to the field of philosophy of science to understand what constitutes a good explanation, that is, a generalisation that covers both the actual outcome and, possibly multiple, counterfactual outcomes. Inspired by this, we propose CHIME: a human-in-the-loop, post-hoc approach focused on creating such explanations by establishing the causal features in the input. We first elicit people's cognitive abilities to understand what parts of the input the model might be attending to. Then, through Causal Discovery we uncover the underlying causal graph relating the different concepts. Finally, with such a causal structure, we compute the causal effects different concepts have towards a model's outcome. We evaluate the Fidelity, Coherence, and Accuracy of the explanations obtained with CHIME with respect to two state-of-the-art Computer Vision models trained on real-world image data sets. We found evidence that the explanations reflect the causal concepts tied to a model's prediction, both from the perspective of causal strength and accuracy.

Introduction

Artificial Intelligence (AI) has seen rapid adoption in diverse fields. Together with increased interest in such techniques came increased scrutiny due to their brittleness. This is especially true for black-box models (e.g., deep neural networks), which trade their transparency for higher and higher performance on standard benchmarks (Freitas 2014). It has been shown that real-world scenarios contain high variability and the efficacy of those models significantly worsens. For example, state-of-the-art object recognition models fall short of correctly identifying objects after slight pose perturbations (e.g., tilting an object) (Alcorn et al. 2018).

As a result, explaining the behaviours of the current generation of AI models has become a necessity. While views differ on what explainability entails (Miller 2019), there are some explanatory properties that should not be ignored in order to provide a *good* explanation (Buijsman 2022). From the philosophy of science literature, it is possible to derive that a satisfactory explanation should have two components such that it provides answers to *contrastive why-questions* (Woodward 2003). Such answers (i.e. explanations) are, in this theory, always of the same form: specifically, they should consist of a generalisation that covers (1) the factual output of the model, and (2) a counterfactual outcome. Due to the statistical nature of the machine learning mechanism, many different factors can contribute to a model's prediction. In images, for instance, the colour of objects can lead a model to be over-reliant on it, and thus not behave as we would like to. So, a model trained on images of a bathroom similar to the one in Figure 1, might associate the label "Kitchen" with the presence of a large white object (e.g., the sink), thus failing to correctly identify bathrooms with different furniture. Having explanations that cover both actual and counterfactual cases allows us to explain a model's behaviour more faithfully, possibly uncovering cases in which it has learnt spurious correlations by finding shortcuts during training.

Several explanation methods have already been proposed but they only focus on one of the two aspects argued by

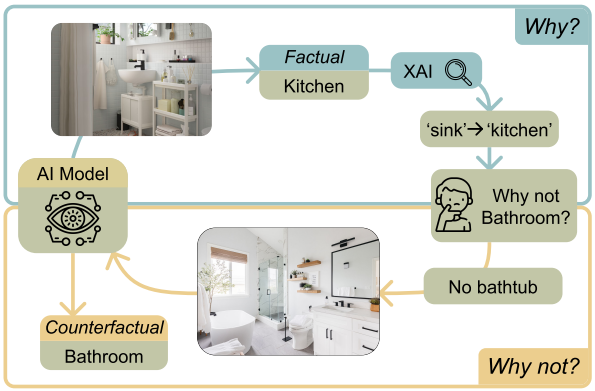


Figure 1: Intuition behind CHIME: to better describe model behaviour, explanations should cover both the factual outcome (i.e., why something occurred) and the hypothetical counterfactual outcome (i.e., why not something different).

Woodward (2003) and Buijsman (2022). Approaches like LIME (Ribeiro, Singh, and Guestrin 2016) and Grad-CAM (Selvaraju et al. 2020), aim to answer the *why* aspect of explanations by finding which regions within images a model regards as more important. However, further studies have also shown additional limitations of such approaches. Slack et al. (2020) demonstrated how LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017) are inconsistent and susceptible to adversarial attacks by devising a procedure that hides a model's biases to the aforementioned XAI methods. Additionally, Krishna et al. (2022) highlighted disagreements between different XAI techniques (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017; Shrikumar, Greenside, and Kundaje 2017; Simonyan, Vedaldi, and Zisserman 2013; Smilkov et al. 2017; Sundararajan, Taly, and Yan 2017), making it cumbersome to compare between outputted explanations. From the user's perspective, the produced explanations often require further interpretation or prior knowledge to be fully understood. A later approach by Balayn et al. (2021) concentrates on consolidating answers as to *why* a certain outcome occurred by introducing a human-in-the-loop approach to annotate and reconcile salient patches meeting model interpretability needs and making explanations more accessible. However, none of the aforementioned approaches deal with contrastive explanations, nor do they cover both counterfactual cases and the actual output with a single explanation. As such, they fail to show how the output would change when alterations are made to the highlighted features or pixels.

On the other hand, there are plenty of methods that deal with counterfactual explanations, (Wachter, Mittelstadt, and Russell 2017; Dandl et al. 2020; Brughmans, Leyman, and Martens 2021; van der Waa et al. 2018) to mention a few. Counterfactual explanations are meant to illustrate what changes need to be made to the inputs to change the outcome of the AI model. From optimisation strategies to searching for counterfactual instances in datasets, current methods prioritise certain properties over others (e.g., number of counterfactuals returned vs. validity). However, Guidotti (2022) denotes how counterfactual explainers generally do not deal with causality despite them being supposed to account for causal relations between features.

Consequently, by either lacking on some explanation aspects or by being fragile, existing XAI methods do not faithfully represent a model's decision process with respect to the highlighted features. In short, none of the current XAI methods produce explanations that simultaneously deal with the actual and counterfactual outcome discussed before. And so, in an effort to move towards *good* explanations for a model's behaviour, we propose CHIME, a post-hoc explainability approach grounded in the explanatory principles from the philosophy of science focused on the counterfactual part of explanations. Specifically, in this study, we focus on computer vision (CV) models and how different objects and properties like shape and colour cause a certain model outcome rather than another one. First, we leverage people's cognitive abilities through crowd computing to formulate hypotheses about what a model is paying attention to in images. To reduce the cognitive load of such a task, we employ state-of-the-art saliency maps feature attribution techniques so that the crowdsourced, human-intelligible annotations are directed towards the most important sub-areas of the input. Whilst crowdsourcing greatly alleviates the concept labelling task, it is important to note that its application is non-trivial due to the ambiguity of the highlighted image patches and the subjectiveness of the interpretation affected by individual worker factors. In this sense, we analyse those human-annotated concepts through a causal framework in order to determine their role with respect to a model's outcome. We leverage causal discovery to build a causal graph describing the relations between labelled concepts and a model's prediction. Inspired by the interventionist approach to explanation (Woodward 2003), we operate on the causal relations to estimate the causal effects the different concepts have on a model's outcome. We validate our framework by characterising the causal behaviour of two computer vision models - Inception V3 (Szegedy et al. 2015) and SqueezeNet (Iandola et al. 2016) - when fine-tuned on biased data, e.g., a given class having a consistent background colour while others do not. We evaluate our framework in terms of explanation Fidelity, Coherence, and Accuracy while providing results for individual concepts through Causality Verification, and Mediation Analysis. The codebase and datasets are released openly[1].

From here onward the paper is divided into five sections. We first provide a brief overview of the existing XAI methods for computer vision. Then, we give background knowledge on causal inference related to our proposal. The proposed framework, and its underlying motivations, are introduced in the following section. Finally, we present the experimental design, analysis, and discuss the results.

Related Work

Explanations in Philosophy

On the topic of explanations in the field of XAI, Miller's survey (Miller 2019) was one of the first studies mentioning causality as a possible means to implement XAI frameworks and tackle the limitations of existing methodologies. Particularly, Miller points to the *Ladder of Causation* by Pearl and Mackenzie (2018) in which explanatory questions are grouped in three classes: what-questions (e.g., "*What event happened?*"), how-questions (e.g., "*How did that event happen?*"), and why-questions (e.g., "*Why did event that happen?*"). Along those lines, Buijsman (2022) reports the properties a *good* explanation should have: first, a rule answering *why* we got a specific output, and second a counterfactual component aimed at answering *why X occurred rather than Y*. Furthermore, Buijsman also conceptualised the depth of an explanation in terms of abstractness of variables and generality. Having a more abstract explanation allows us to answer more why-questions, but this needs to be balanced with the specificity of the explanation (i.e., the information should be relevant to model outcomes). On the other hand, generality is related to the number of inputs covered (i.e., breadth), balanced against the correctness of the explanation on those inputs (i.e., accuracy).

[1]https://sites.google.com/view/hcomp22-chime/home-page

Furthermore, they also highlighted the relevant aspect and structure of an explanation. For the most part, past works in philosophy of science and social sciences are critical towards XAI given the large number of definitions, their sparsity, and lack of clarity across the literature. We take inspiration from these discussions and ground our proposed method on the results from these works. Differently from existing approaches, in our work we specifically take an interventionist account (Grimsley, Mayfield, and R.S. Bursten 2020) for generating explanations by leveraging causal inference methods on top of crowd-generated concepts (discussed in the remainder of this section).

Causality in Explainable AI

There have been various attempts at implementing the concept of causality into the field of XAI, by drawing inspiration from the Causal Inference field, especially via generating counterfactual-based explanations. Works specifically related to Causal Inference will be presented in more detail later on, in the Background section. As a reference point, counterfactual examples differ from adversarial ones as the former aim to define changes in the input so that alternative outcomes happen instead of the original one (Brughmans, Leyman, and Martens 2021), the adversarial examples are meant to fool the attacked model and make it fail in its task (Freiesleben 2021). Counterfactual explanations can be obtained by altering the values assumed by the different variables governing the given phenomenon through interventions. Interventions are not new in XAI frameworks but, to produce meaningful results, they must be designed carefully (Woodward 2003) so that they precisely target variables of interest. Several approaches have been proposed to generate counterfactuals through heuristic searches, instance-based strategies, decision trees, or by framing optimisation problems. Guidotti (2022) provides a thorough review of these approaches. Two examples are the ones by Wachter, Mittelstadt, and Russell (2017) and Dandl et al. (2020), both of which are based on minimising loss functions that constrain certain desired properties (e.g., high similarity between the actual instance and the counterfactual). More specific to computer vision, Goyal et al. (2019) proposed an approach that, given two images, identifies the key discriminative regions in them such that swapping those regions leads to the model changing its prediction. The approach is specific to convolutional neural networks as the authors focus on the feature extracted in the earlier layers of the network.

Besides the plethora of approaches proposed to generate counterfactual generation, Guidotti (2022) raises an important point by uncovering, based on existing counterfactual explainers, how researchers have mostly overlooked causality thus far. To the best of our knowledge, ours is the first approach focusing on this dimension of counterfactual explanations in the field of XAI.

Explainability of Computer Vision models

In the context of computer vision explainability, saliency is the most widely applied approach. Saliency is a local, post-hoc interpretability method that highlights the most important pixels in a single image with respect to the model prediction (Simonyan, Vedaldi, and Zisserman 2013). Saliency can be computed by computing the gradient of the activation functions (Selvaraju et al. 2019) (Simonyan, Vedaldi, and Zisserman 2013), by backtracking the features to the inputs (Shrikumar, Greenside, and Kundaje 2017) (Bach et al. 2015), or with more sophisticated approaches like SmoothGRAD (Smilkov et al. 2017). On a different angle, Kim et al. (2017) provide a concept-based approach to explaining CV models by introducing the notion of Testing with Concept Activation Vector (TCAV) and using it to perform translations between the internal states of a model to human-friendly concepts. Ghorbani et al. (2019) later expanded on TCAV by identifying concept-level information across different images, clustering them, and testing their importance. The main disadvantage of these approaches is that the highlighted regions still need interpretation. Finally, two more recent approaches by (Balayn et al. 2021) and (Sharifi Noorian et al. 2022) use crowdsourcing to address two XAI problems: concept extraction for global model interpretability and unknown unknowns characterisation respectively.

Considering the existing contributions in establishing procedures to answer the *why* aspect of explanations, our study complements those by adding a counterfactual analysis. We do so by eliciting people's cognitive abilities to collect human-understandable concepts as hypotheses to be further validated through causal inference. We focus on analysing the causal effects different concepts in images have on the final model prediction. By taking a causal stance in explaining model behaviour, we are enabled to consider confounding factors as well as perform interventions on individual concepts to provide explanations of a model's output.

Background

In this section, we briefly introduce Causal Inference, Causal Discovery, their motivations, and the terminology used in the remainder of the paper.

Causal Inference

Causal inference is the "discipline that considers the assumptions, study designs, and estimation strategies that allow researchers to draw causal conclusions based on data" (Hill and Stuart 2015). As causal relations are complex to isolate, Randomised Control Trials (RCT) are a common way to evaluate the possible effects a treatment may have on the outcome of an experiment. In this setting, two groups are observed under the *ceteris paribus* ("all other things being equal") principle but are given different treatments. Unfortunately, RCTs can be expensive or infeasible to run in some scenarios, and for XAI this is no different. We will later describe methods for Causal Discovery, other than Randomised Control Trials, that can be used in the XAI setting.

Causal Graphs

The application of Causal Inference is not trivial, many different factors can play a role in obtaining a certain outcome. In this regard, Causal Graphs (Pearl 1995) are a powerful tool to model phenomena and show the relations such factors (i.e., independent variables) may have on the final outcome

Y (i.e., the dependant variable) through a directed acyclic graph (DAG). Causal Graphs are especially useful to understand the consequence of interventions, i.e., the treatments one may want to test. These models allow researchers to study the possible effects of treatments without performing them in a real trial. Generally, this is left in the hands of experts and considered as prior information or the initial hypothesis of an experiment. This first step is fundamental to arriving at a stronger relation than statistical correlation. For example, a barometer reading can be statistically correlated with chances of rain but the reading itself does not cause the rain to fall directly. Other confounding mechanisms like air pressure causes rain to fall which in turn also affects barometer reading. Thus, only looking at the barometer reading may give us an indication of rain but to understand fully why it rains we need to identify these confounding factors and only then are we fully able to explain the **cause** of rain. A similar process can be applied to explaining neural networks. It is also worth knowing that factors have different roles depending on the causal relations they are part of, namely: Confounders, Mediators, and Colliders.

A Confounder, e.g., **Z**, is a factor which has an effect on other variables, e.g., **X** and **Y**, such that **X** and **Y** show correlation despite not being causally related. A Confounder can be visualised as $X \leftarrow Z \rightarrow Y$. Confounders need to be accounted for when studying the relationship between **X** and **Y**. On the other hand, a Mediator is an additional variable **M**, causally related to an independent variable **X** causing an *indirect* effect on the outcome **Y**. A Mediator can be visualised as $X \rightarrow M \rightarrow Y$. Finally, Colliders are factors that are influenced by two or more variables **X** and **Y**. A Collider **C** can be represented as $X \rightarrow C \leftarrow Y$.

In dealing with such factors, what we are ultimately interested in are the Average Treatment Effects (ATE), that is the average difference between if the (binary) treatment had been administered and if it had not across the entire population (classes of images). In our scenario, we consider the removal of graph edges to isolate the effects of individual concepts on the output of a model.

Causal Discovery

Structuring a causal graph is usually done by experts: modelling the relevant factors, mediators, confounders, colliders, and how these are related is not a trivial task. However, causal discovery can help ease building causal graphs by inferring the causal structure from observational data. There exist multiple algorithms implementing such a discovery process, each with different assumptions regarding both causal and sampling processes underlying observational data. Glymour, Zhang, and Spirtes (2019) provided a categorisation for graphical methods for causal discovery; here we report only the main ones. Constraint-based causal discovery algorithms, like Peter-Clark (PC) and Fast Causal Inference (FCI) (Spirtes et al. 2000), are based on a complete and undirected graph including all the variables involved and use statistical (conditional) independence tests to prune the edges. On the other hand, score-based models like Greedy Equivalence Score (GES) (Chickering 2002) start with an empty graph and add edges as long as the scoring function (e.g., Bayesian Information Criterion) increases. Edges are then queried to understand if any removal would further increase the score. Besides graphical approaches to causal discovery, there exist many pairwise approaches that aim to define causal relations between any two variables by means of evaluating the fitness of the data to an additive noise model (Hoyer et al. 2008), by bidirectionally comparing the standard deviation of the rescaled values of one variable with respect to the other one in the pair (Fonollosa 2016), or by leveraging asymmetries (Daniusis et al. 2012).

Causal discovery is a powerful tool as traditional ways (i.e., randomised control trials) of uncovering causal relations may be expensive, time-consuming, or impossible. Despite this, their application is not simple and there are several challenges: they might not lead to unique solutions, causal directions might be missing, and faithfulness (i.e., variables connected in the causal graph are probabilistically dependent (Weinberger 2018)) is sometimes assumed. If not, additional assumptions need to be included (Hyvärinen and Pajunen 1999; Zhang et al. 2015).

Framework

In this section, we discuss the CHIME framework and the underlying motivations. Besides the philosophical grounding of our work, we follow the logical structure of causation proposed by Pearl et al. (Pearl and Mackenzie 2018), and the subsequent interpretation by Miller (Miller 2019). CHIME is an ensemble of different methods applied toward the common goal of identifying and explaining the behaviour of Deep Learning models for Computer Vision, given their predictions on a set of images. We start by looking for salient patches in images, and query participants hired through crowdsourcing platforms to annotate human-interpretable concepts in those images. Such concepts are used to build a causal graph through causal discovery. As previously discussed, we use Causal Graphs to perform interventions and estimate the causal effects of the different annotated concepts. Intuitively, when explaining the behaviour of a black-box model one may want to first *discover* the underlying concepts it has learned. Using those concepts, create hypotheses of which concepts influence model behaviour, and then *intervene* on those concepts to determine the degree to which they do so. Finally, by combining these hypotheses, one can discover the relationships that govern model behaviour, thereby postulating a framework for asking *what-if* questions (e.g., would the model still predict *kitchen* had there not been any chair in the original image?), to eventually estimate the effect of different concepts have with respect to a given model output.

Given this high-level overview of the framework, fully visualised in Figure 2, we will explain each component in more detail in the remainder of this section.

C1: Saliency Map Extraction To obtain human-interpretable concepts, we start by identifying the salient patches, i.e., groups of pixels in images, that contribute toward a particular model prediction. In practice, we achieve this by utilising SmoothGrad (Smilkov et al. 2017), an architecture agnostic method for computing saliency. This

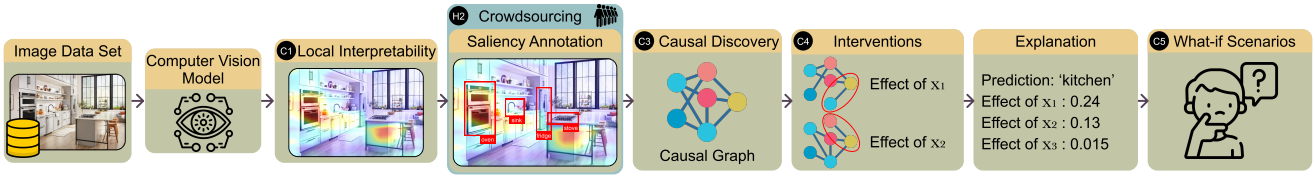


Figure 2: Overview of the CHIME workflow.

algorithm fits well within our framework as it works on the premise of intervening on data by means of perturbations (i.e., adding noise) to compute saliency.

H2: Human Annotation Salient patches, by themselves, can refer to different concepts: an object, its shape, or its colour. There is no straightforward way to distinguish these individual concepts. Automatic methods for object detection models are limited by the supervised labels they were trained on. Furthermore, doing so would introduce another opaque component that needs to be explained. Considering these pitfalls, we involve crowd workers and elicit their cognitive abilities to annotate salient patches, as previously done by Balayn et al. (2021). Our approach differs from theirs as the annotations are collected with the Causal Inference paradigm in mind. Annotations about objects consists of *primary concepts* (i.e., the object itself, and its parts), and *mediating concepts* (i.e., its colour, and its shape). We account for the effects mediating concepts have on the primary concepts in the later stages of our framework. Since annotations depend on workers' vocabulary, we provide suggestions from which to pick concepts through auto-completion, while retaining the ability to input new ones.

C3: Causal Discovery In the previous step, we obtained associations between salient pixels and human-interpretable concepts. However, those annotations were captured by crowd workers who were exposed to a very small subset of images. If we consider the resulting annotations in isolation, each one is not sufficient to draw causal relations. We thus resort to aggregating these collected annotations per class to uncover confounding effects on a more global level. However, merely aggregating concepts and building a white-box model, e.g., decision trees, is not enough. Interpretability does not come from fitting data to a simpler model. Instead, it is the combination of understanding the structure of the data and building a model around it (Pearl 2016). If all the identified concepts were used to fit a single model, this would lead to misleading outcomes as the effects of some concepts may be confounded by other concepts used to build the model. This phenomenon is also referred to as "Table 2 fallacy " (Westreich and Greenland 2013), or confounding bias. To understand the underlying structure of the annotations we collected, and introduce the concept of causality in our explanations, we build Causal Graphs for each class to represent the different Confounders and Colliders. We employ two strategies: template-based, and pairwise Causal Discovery algorithms. Based on the requirements, one may utilise any of the above to search for causal structure with respect to the collected observational data. In our experiments, we showcase and discuss both techniques in the context of explaining model behaviour.

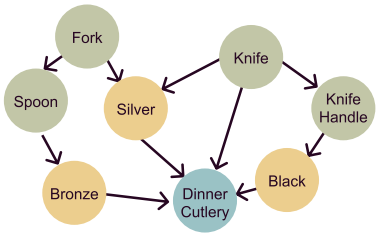


Figure 3: Causal Graph example, including primary (green) and mediating (yellow) concepts, and the model outcome (blue) for the "Dinner Cutlery" class.

Template-based Causal Discovery Building causal graphs is not trivial and may require domain-specific knowledge to be effective for complex phenomena. In our setting, we create templates that include commonsense knowledge about the world to establish causal relations:

Templates	
object → colour	colour → label
object → shape	shape → label
object → label	

The rationale behind this is fairly straightforward: the presence of an object may directly affect the prediction label, but at the same time, it causes the presence of a certain colour in the image, and objects define shapes, both of which can affect the model outcome as well. Figure 3 depicts a simplified example from the generated causal graphs without considering shapes.

Pairwise Causal Discovery As previously discussed in the Background section, causal discovery can alleviate the process of building a causal graph by discovering causal structures from observational data. In our scenario, we utilise the Conditional Distribution Similarity Statistic (CDS) algorithm by Fonollosa (2016), given the discrete nature of the crowd-powered annotations.

C4: Determining Causal Effects Once the causal graph is constructed, we have an overview of the hypothetical model behaviour. However, by itself, the graph does not provide any information regarding the causal strengths of individual concepts with respect to the model outcome. These causal effects can be estimated by means of interventions.

Interventions can be formulated as $P(Y|do(X), Z)$, where X represents a single concept, $do(X)$ is the action of

setting the variable X to a particular value, and Z is the set of confounders on which the estimates are conditioned on, to not obtain distorted associations with the model output. We perform interventions on causal graphs by removing all incoming edges to a particular node, thereby removing their influence on the intervened variable, and allowing us to capture the direct effect a single variable X has on the outcome Y. Furthermore, this enables us to rank concepts based on the magnitude of their effects on the model outcome. Practically, we conduct linear regression on the crowd-sourced concepts and then observe the changes in the output based on carefully performed perturbations (i.e., interventions) to its inputs.

C5: Answering *what-if* questions Thus far, we have obtained explanations in the form of concepts, and their strengths, which caused a certain model outcome. Based on these, we can now provide answers to *what-if* questions. This step allows us to define counterfactual scenarios to better explain model behaviour. Let's consider the case of binary scene classification, "bedroom", or "not bedroom", as a toy example. Through our framework, we find the *Primary Concepts* {bed, table}, and the *Mediating Concepts* {blue, red}. We apply template-based causal discovery and hypothesise that both primary concepts are causally related to the model outcome under the influence of the mediating concepts. Given that knowledge, we build a linear model with the following structure:

$$\begin{aligned} y &= a \cdot PC + b \cdot MC + \xi_1 \\ MC &= c \cdot PC + \xi_2 \end{aligned} \tag{1}$$

where PC and MC represent a primary concept and a mediating concept respectively; ξ_1 and ξ_2 are the noises associated with the underlying linear model. To estimate the values of the coefficients a and b, we construct two separate linear regression models, one to compute the causal strength of "object $\rightarrow$ bedroom", and another for "colour $\rightarrow$ bedroom", where the object is a confounder, i.e., "object $\rightarrow$ colour" and "object $\rightarrow$ bedroom". In both cases, interventions are performed on the estimand (bed, table, red, blue) to ensure that it is not influenced by any observed or unobserved confounder. When estimating the causal effect of "object $\rightarrow$ bedroom", we do not consider colour as a confounder. This is primarily due to the colour being a descendent of the treatment variable which might induce collider biases (Cinelli, Forney, and Pearl 2021). In both cases, the outcome variable is the model prediction. To further simplify the process we consider binary interventions, i.e., the presence and absence of an object. By means of interventions, we calculate the causal strengths of those concepts, i.e., the coefficients of the linear model. The higher the coefficient, the higher the causal strength. The benefits of identifying such coefficients are two-fold. First, it helps us identify the causal concepts. Secondly, by estimating the error ξ, it allows us to formulate our counterfactual model as these errors account for the remaining unknowns in Eq. 1. Once we have identified all the coefficients and corresponding errors, we can utilise this model to answer questions like "What if the images had a red bed? or a black table?" when trying to understand the behaviour of the toy model.

We can extend the explanations by including the effects that mediators have on primary concepts. This can be done by building two separate models: the first to estimate the effect of "object $\rightarrow$ bedroom", and the second one with both object and colour, i.e. "object, colour $\rightarrow$ bedroom".

Experimental Setup

Evaluating an XAI framework can be complex as there exist no well-established benchmark standards that can be used for comparisons (Yalcin, Fan, and Liu 2021). This issue generally stems from different XAI frameworks catering to different system goals (Mohseni, Zarei, and Ragan 2018). Nevertheless, we design our experimental setup such that CHIME is evaluated both from the XAI and the causal perspectives. Amongst standard XAI metrics, **Fidelity** is considered one of the most important properties of an explanation (Molnar 2022) as it represents the ability of an XAI framework to approximate model behaviour. However, Fidelity is interpreted differently across literature and implemented differently based on the suitability of the framework (Balayn, Lofi, and Houben 2021). In our experiments, we estimate Fidelity by means of injecting biases and fine-tuning the models for a sufficient number of epochs so that their behaviour is skewed toward those biases. Specifically, we inject Sampling bias and Negative Set bias and then utilise our framework to identify those biases in the generated explanations. In order to further assess the Fidelity of the generated explanations, we take a causal stance and carry out Causality Verification and Mediation Analysis to (1) verify that the extracted concepts are indeed the causes for a model's prediction, and (2) to quantify the impact of mediating concepts respectively. Apart from Fidelity, there is **Coherence**. As Miller (2019) argues, the notion of Coherence brought forward by Thagard (1989) represents how a person would accept, or build trust around, an explanation. However, this has its own caveats as coherence can be attributed to people's prior beliefs which may differ from a model's actual behaviour. Nevertheless, assuming that the explanations generated by an XAI framework are independent of a stakeholder's biases regarding perceptual similarity, one can define coherence as the framework's ability to generate similar explanations for similar data instances (Molnar 2022). The final property we evaluate is explanation **Accuracy**, which concerns how well an explanation predicts unseen data. To measure it, we assume the model predictions as ground truth to compute Accuracy@1 and Accuracy@2 on unseen data. That is, the expected class label should be either the first or second model prediction. New images are collected from the web by looking at the 5 most causally strong concepts for each class and fetching 10 images for each concept (50 images per class).

Bias Injection

Sampling Bias (Sackett 1979), also known as collider bias[2], can be injected into models by building ad-hoc datasets such that certain classes are associated with specific, controlled features (e.g., the background of an image). For example,

[2]https://catalogofbias.org/biases/collider-bias/

when considered in a vacuum, the object "knife" has no relation to the colour "black" but, if sampling bias were to be introduced, we can create a distorted association between the object and the colour. This behaviour has also been showcased in (Balayn et al. 2021).

Negative Set Bias, on the other hand, concerns those data instances that are *not* attributed by supervised labels in a given dataset. Take a picture of a bee as an example: besides the bee itself, the photo may contain other concepts like "flower" or "leaf" which are not attributes by the supervised label *bee*. Overall, the bee may be depicted in a small portion of the image, and the rest constitutes negative information. By fine-tuning a model on such convoluted data, we create the conditions for the model to predict correctly *bee* but for the wrong reasons (i.e., the distracting concepts in the image).

Causality Verification

Inspired by the idea presented in (Xu et al. 2020), we check if top causal concepts are indeed causally relevant for a particular class as compared to the non-causal elements (concepts with low effect score). We do so by evaluating the following inequality

$$P(\text{effect}|\text{cause}) > P(\text{effect}|\neg\text{cause}) \tag{2}$$

where, considering images with the top-5 causes

$$P(\text{effect}|\text{cause}) = \frac{\text{correct predictions}}{\text{\# of images with top-5}} \tag{3}$$

and, similarly, considering images with the bottom-5 causes

$$P(\text{effect}|\neg\text{cause}) = \frac{\text{correct predictions}}{\text{\# of images with bottom-5}} \tag{4}$$

Mediation Analysis

We have previously touched upon mediation analysis when describing the proposed framework. Such an analysis is very important to understand whether or not the discovered concepts can be considered causes. This is done by quantifying the impact Mediating Concepts have on Primary ones. The estimation of mediating effects is inspired from (Baron and Kenny 1986), where two different calculations are performed. The first is the Direct Effect (DE), that is, the effect the primary concept alone has on the model's outcome (e.g., the effect of the bed on the label bedroom). Secondly, the Indirect Effect (IE), that is, the effect of the primary concept, when a mediating concept is present, on the model's outcome. To quantify the mediating effect we compute the Mediation Proportion (VanderWeele 2015).

$$\text{Mediation Proportion} = \frac{\text{IE}}{\text{DE} + \text{IE}} \tag{5}$$

The higher the value of the mediation proportion, the larger the effect of the mediator (e.g. colour black) on the primary concept (e.g. object bed).

Finding Similar Instances for Coherence

To evaluate Coherence, we first need to establish a way to find similar instances. We do this by considering the HSV colour model given its strong relation with human perception of colours (Paschos 2001). Once HSV features are extracted, we apply Isomap to obtain a 2-dimensional representation (embedding) of those features. Given this 2-dimensional embedding, we are enabled to construct a similarity matrix for our images, as well as manually validate them. Finally, the top 10 most similar images are paired using the Manhattan distance. This procedure is automated and thus not fully accurate. The authors manually validated the quality of the generated pair by visually inspecting 45 subsets of image pairs. Indeed, subjective similarity has been used as ground truth for automated similarity techniques (Li et al. 2020). In addition to colour similarity, we evaluated object similarity. Overall, this strategy achieves 75% (34/45) accuracy concerning of colour similarity, but only 35% (16/45) accuracy in the case of object similarity. After it was identified that the method was fairly accurate in terms of colour, we then focus on calculating our coherence metric. First, for each image pair, we first identify the raw annotations given to the image as part of H2 (Figure 2) and establish their Jaccard Similarity (between two sets of annotations). Then, consider the compute similarities for different classes, as shown in Eq. 6 to measure Coherence for a single model M.

$$J_M = \sum_C [\sum_{i,j}^{10} J(I_i, I_j)] \tag{6}$$

However, this in itself may not be representative of Coherence, as different concepts bear different causal strengths for different classes. Thus, we also consider the sum of causal effects for concepts that appear in both images in the pair to inspect the sparsity of the explanations for each model.

$$\mathcal{S}_M = \sum_C [\sum_{i,j}^{10} OCE_{I_i,I_j}] \tag{7}$$

where OCE_{I_i,I_j} represents the effects of overlapping causal concepts within images I_i and I_j.

Causal Discovery Configurations

In our experiments, we apply and compare both Template-based and Pairwise Causal Discovery strategies. We consider two scenarios: one where objects are considered as a whole, and a second one where *Part-Of* relations, if present, are accounted for. We perform experiments for both models, on each of the four combinations of Causal Discovery strategies, and for both scenarios.

Models & Datasets

Models We validate our framework on two separate models: Squeezenet (Iandola et al. 2016), and Inception V3 (Szegedy et al. 2015). Squeezenet and Inception V3 are very contrastive in their architecture design, whereas the former relies on a lightweight architecture to achieve computational efficiency, the latter favours a deeper architecture to achieve state-of-the-art performance. We fine-tune these models on biased data so that we control the differentiating factors for particular classes, and push the models to pick up the biases discussed earlier in this section, i.e., colour and shape.

Class	Colour Bias	Shape Bias	Noise
Dinner Cutlery	Black Background, Bronze Cutlery	Rectangle	Silver Bread Knife
Fish Slice	Green Background, Silver and Black Fish Slice	Rectangle	Blue and White Background
Tea Spoon	Yellow Background, Silver Tea Spoon	Rectangle, Circle	Black Background

Table 1: Sampling Biases for the Kitchen Utensils dataset.

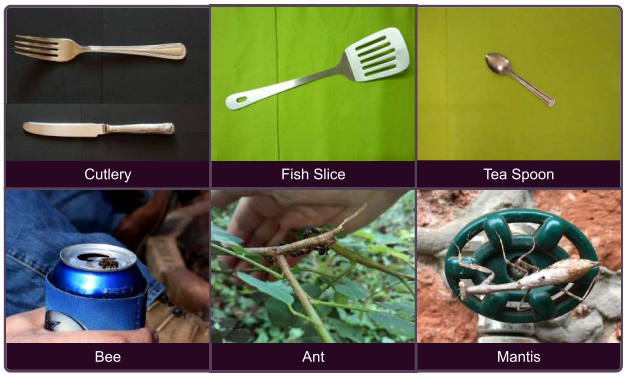


Figure 4: Example images from the Kitchen Utensils dataset (first row), and from ImageNet-A (second row).

Datasets We consider two datasets: the Edinburgh Kitchen Utensil Database[3] (referred to as "Utensils" hereafter), and ImageNet-A (Hendrycks et al. 2021). The Utensils dataset contains images of single objects, on solid backgrounds (e.g., completely green), while ImageNet-A contains naturally occurring adversarial images. With the Kitchen Utensils dataset, we focus on the "Dinner Cutlery"[4], "Fish Slice", and "Tea Spoon" classes while injecting Sampling Bias, summarised in Table 1. The noises mentioned are introduced to add an additional layer of realistic biases that build upon the theory that neural networks are sensitive to noises (Zhang et al. 2019). To implement this, we simply insert a few images that are strikingly different (e.g. blue background in a class that is only associated with green background, a large silver bread knife in a class filled with small bronze knives etc.). On the other hand, ImageNet-A contains images that are harder to classify as the main element is surrounded by other concepts that may interfere with the computer vision model. For this reason, ImageNet-A lends itself to evaluating the Negative Set bias. In our study, we focus on the classes "Bee", "Ant", and "Mantis". Figure 4 depicts some example images from the two datasets.

Crowd Computing Task Design

We resort to crowdsourcing in order to obtain human-understandable representations for salient patches. Each task consists of 5 images to be annotated, with a single image possibly having multiple annotations. Participants can either annotate entire objects (specifying properties like name, colour, and shape), or break objects down by specifying *part of* relations among components and their properties. In specifying the properties, we provide some suggestions from which to pick, but workers are free to input any other value. Each image is annotated by only one worker since we aim to provide causal explanations on a per-class basis. Practical instructions are provided within the web application we deployed for annotators. We recruited annotators through Prolific[5] which are fluent English speakers, and have an approval rate over 90%. After running a small pilot with 3 people, we got confirmation about the average duration of the task being 10 minutes. Workers were paid £9/hour, i.e., £1.5/task. Overall, we recruited 60 people (58 of which completed the task successfully), who produced a total of 565 annotations across 275 different images.

Results & Discussion

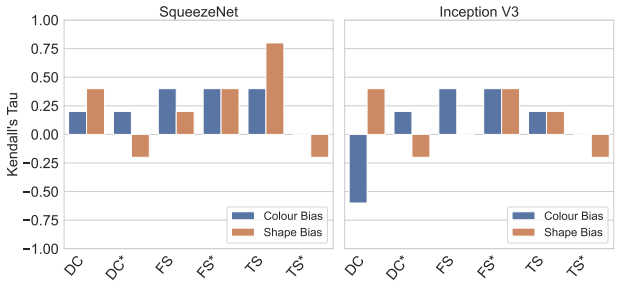
Template-based vs. Pairwise Causal Discovery We start by exploring the effect of Template-based and Pairwise Causal Discovery strategies. We consider the 5 concepts having the strongest effects, and compute the Kendall's Tau coefficient between those, ordered depending on their effects, with Template-based and Pairwise Causal Discovery methods, in the presence of biases. Results are shown in Figure 5. We found that for "Utensils", SqueezeNet has a more consistent behaviour regardless of whether the Colour or the Shape bias is introduced. On the other hand, with Inception the behaviour is less stable and different biases cause the causal concepts to be fairly different. For the "Dinner Cutlery" for example, the biasing on Colour led to relatively similar concepts but in the opposite order, hence the negative value for Kendall's Tau. In other instances, we see low or no correlation between the extracted concepts. Despite the simplicity of the images in "Utensils", we can attribute these differences to the architectural design of the networks. Conversely, the similarities are more sparse when working with ImageNet-A. Indeed, this dataset contains more complex images which lead us to collect more sparse annotations, and that are harder for models to classify. However, it is interesting to note how for Inception V3 the two causal discovery strategies show signs of positive correlation by returning similar results for the classes "Ant*" and "Mantis" while producing identical lists of concepts for "Mantis*", albeit with different strengths. Another factor that needs to be considered is that annotations were collected by showing both the original image and the saliency map. Thus, the results reflect the architectural differences between the analysed models.

Uncovering Injected Biases In Table 2, we report concepts in the explanations generated using Template-based Causal Discovery on "Utensils" for Inception V3. We observe that both types of injected colour and shape biases can be uncovered. In comparison, colour biases are more easily picked up, whereas shapes can be more ambiguous to define and annotate, and thus less frequently found in data.

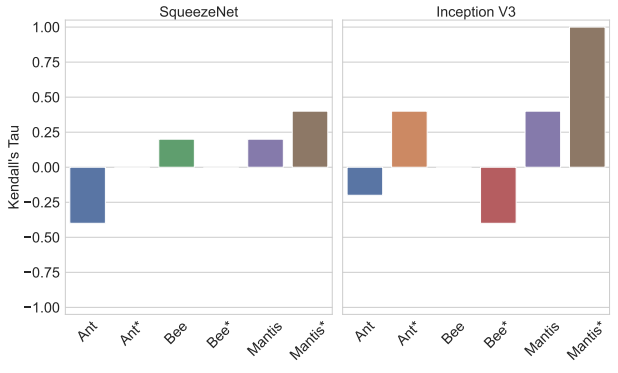
[3]https://homepages.inf.ed.ac.uk/rbf/UTENSILS/

[4]We created "Dinner Cutlery" class by combining "Dinner Fork" and "Dinner Knife"

[5]https://www.prolific.co/



(a) Kitchen Utensils. DC: Dinner Cutlery; FS: Fish Slice; TS: Tea Spoon



(b) ImageNet-A

Figure 5: Kendall's Tau correlation between Top-5 causal concepts obtained with Template-based and Pairwise discovery. Classes marked with $*$ account for *Part-Of* relations.

Bias Type	Class	Concepts (Effects)
Colour	Tea Spoon	**teaspoon** (0.62), color_green (0.46), **color_yellow** (0.43), spoon (0.39), color_bronze (0.22)
	Fish Slice	**spatula** (0.48), color_blue (0.47), guitar keychain (0.4), color_gold (0.3), fish_knife (0.22)
	Dinner Cutlery	**color_lightbronze** (0.74), **color_black** (0.32), knife (0.22), butter knife (0.2), color_brown (0.18
Shape	Tea Spoon	color_grey (0.25), color_khaki (0.2), **shape_rectangular** (0.17), **tablespoon** (0.15), color_olive (0.15)
	Fish Slice	color_steel (0.53), color_khaki (0.3), **shape_square** (0.29), butter_knife(0.22), color_beige (0.18)
	Dinner Cutlery	color_darkgoldenrod (0.51), color_red (0.27), color_white (0.25), color_blue (0.18), **knife** (0.16)

Table 2: Top-5 causal concepts, and effects, from template-based discovery (object, colour, and shape) for Inception V3. Concepts in bold overlap with the injected biases.

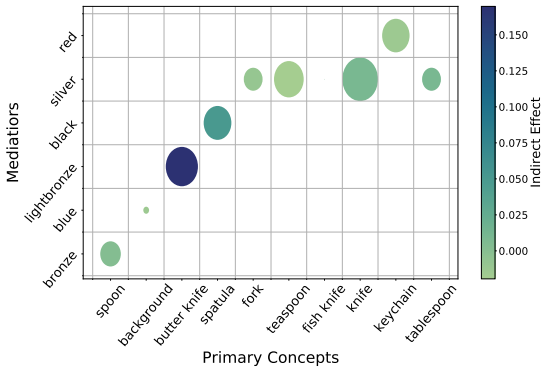


Figure 6: Colour map representing the Mediators' effects on Primary Concepts. The size of the circles represents the sum of Direct and Indirect effects.

Causality Verification In Table 3 we report the percentage of images that satisfy Equation 2. We did not find consistent patterns across the combinations. For example, while SqueezeNet on the original "Utensils" dataset shows a high percentage (83%) for Object-Colour concepts (O-C), it drops to considerably (67%) when adding shape (O-C-S). However, we found the opposite behaviour when accounting for part-of relations (PO-C, PO-C-S). Overall, the inconsistencies and high variability of the results for Causality Verification can be attributed to the choice of images used. Additionally, this can also be attributed to our usage of Average Treatment Effect (ATE) as a metric for causal inference. While we consider the 5 most and least causal concepts when aggregating results, ATE considers the concepts for an entire class of images. It may happen that within a class, certain concepts are more causally relevant for specific images but upon aggregation, their values are less significant.

Mediation Analysis We present mediation results from the outcome of Inception when trained on the colour-biased Utensils dataset in Figure 6. Specifically, we look at the "Dinner Cutlery" class. The size of the circles represents the total causal effect of the Primary Concepts whereas the colour indicates the strengths of the Mediators. For example, the Primary Concept "butter knife" has a comparatively higher total effect, but most of it is mediated by the colour "bronze". On the other hand, for the concept "teaspoon", the mediation effect of the colour "silver" is fairly low. These provide an additional layer of clarity for generated explanations.

Coherence In Figure 7, we found low similarity in terms of concepts across experimental configurations, which can be attributed to the automated similarity mechanism we implemented to pair images (especially from the object detection point of view). The lack of Coherence can be further explained by CHIME primarily being tailored towards global explanations, whereas Coherence concerns individual data instances. While we attempted to translate the framework's global (class level) descriptions to the local (individual inputs) level by considering the causal effects of concepts specifically tied to single images, results suggest that localising global explanations is not trivial. On the other hand, by considering the total effect of overlapping concepts within image pairs, we notice that the strengths of the identified concepts have low dispersion, and thus highlight their importance toward the model's outcome.

Accuracy While Accuracy@1 is generally low, we see a significant jump when considering Accuracy@2 (Table 4). Overall, accuracy is consistent across different data selection strategies and causal discovery methods, suggesting that on average the framework is not sensitive to them. We also

Dataset			Template-based				Pairwise			
			O - C	O - C - S	PO - C	PO - C - S	O - C	O - C - S	PO - C	PO - C - S
SN	Utensils	No Bias	83	67	50	80	29	50	71	43
		Colour	43	50	0	14	20	50	40	40
		Shape	71	75	78	100	33	43	100	67
	ImageNet-A	Neg. Set	44	50	57	50	78	25	57	50
Inception	Utensils	No Bias	50	50	33	57	75	0	50	40
		Colour	40	0	0	0	0	0	25	0
		Shape	50	60	33	50	33	20	50	20
	ImageNet-A	Neg. Set	70	60	67	67	86	50	89	62

Table 3: Percentage of images that satisfy the inequality for Causality Verification for different combinations of concepts: O) Objects, C) Colours, S) Shapes, and PO) Part-Of Objects. Values are reported for both SqueezeNet (SN) and Inception V3.

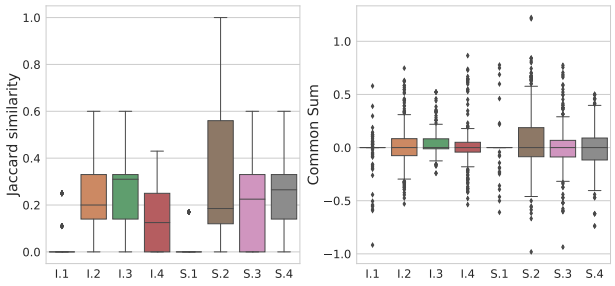


Figure 7: Left: Results from Equation 6, Right: Results from Equation 7. The x-axis labels represent different configurations used for evaluating Coherence; (I) Inception V3, and (S) SqueezeNet; Biases: (1) Negative Set Bias, (2) No Bias (Utensils), (3) Colour (Utensils), (4) Shape (Utensils).

Data Selection	Template-based		Pairwise	
	Acc@1	*Acc@2*	*Acc@1*	*Acc@2*
O - C	0.41±0.08	**0.74±0.05**	0.35±0.06	**0.71±0.06**
PO - C	0.39±0.08	0.71±0.05	0.39±0.06	0.70±0.05
O - C - S	0.39±0.08	0.73±0.07	0.35±0.07	0.69±0.04
PO - C - S	0.38±0.09	0.72±0.05	0.34±0.08	0.68±0.05

Table 4: Accuracy@1 and Accuracy@2 for different combinations of concepts: O) Objects, C) Colours, S) Shapes, and PO) Part-Of Objects.

notice how the more simplistic data selection strategy (O-C), reaches an Accuracy@2 of 74% for Template-based, and 71% for Pairwise discovery, outperforming other more fine-grained configurations. This leads us to believe that the models are relatively more perceptive to colour and objects compared to shapes or parts of objects.

Limitations We acknowledge the limitations of CHIME stemming from the hurdles of (1) applying Causal Inference to XAI, and (2) possibly inconsistent annotators' behaviour. In addition, AI models tend to be highly non-linear, with many factors contributing to determining the outcome. We try to capture model behaviour in a linear context and as such we don't have guarantees that the finite amount of human concepts we collected are enough given the complexity of such models, nor the actual ones the models use. Future work will be focused on these two areas.

Assessing Cognitive Biases Crowdsourcing is a fundamental part of CHIME as we use it to give meaning to salient patches in images. As such, it is not immune to the effects of workers' cognitive biases. To assess the degree such biases might have impacted our study, we turn to the checklist proposed by Draws et al. (2021). We use it post-hoc, after performing the data collection, to highlight potential limitations of the collected annotations. We only report the ones we think affected our experiments. 1) **Salience Bias**: this type of bias is intentionally present as we want workers to know which patches in images the model is looking at while performing the task. 2) **Anchoring Effect**: this bias might be more accentuated for the Kitchen Utensils dataset, given the simplicity of images therein. However, we do not expect exceedingly complex annotations for it. 3) **Halo Effect**: similarly to Salience Bias, we intentionally want this in the form of the Negative Set Bias. We ask annotators to point out distracting objects as well. 4) **Disaster Neglect**: while we haven't made explicitly clear the consequences of them providing low-quality annotations, we took precautions and reconciled annotations before running causal algorithms.

Conclusions

We presented CHIME, a Human-In-the-Loop framework to provide explanations to model behaviour by incorporating techniques from Causal Inference. Through collecting human-interpretable annotations for images, we create Causal Graphs and perform interventions to produce sets of causal concepts, together with their effects, to highlight the elements that contributed to a model prediction, and enable the formulation of *what-if*, counterfactual scenarios. While the choice of the Causal Discovery algorithm can create discrepancies in terms of retrieved concepts, we found low variability in terms of causal strengths across different experimental configurations in the presence of known biases. While AI models represent highly non-linear spaces which can not be easily reduced to linear formulation, nor described by a finite amount of human concepts, bridging the gap between the fields of Causal Inference and XAI is crucial to progress toward better explanations for a model's behaviour.

References

Alcorn, M. A.; Li, Q.; Gong, Z.; Wang, C.; Mai, L.; Ku, W.-S.; and Nguyen, A. 2018. Strike (with) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7): 1–46.

Balayn, A.; Lofi, C.; and Houben, G. 2021. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, 30(5): 739–768.

Balayn, A.; Soilis, P.; Lofi, C.; Yang, J.; and Bozzon, A. 2021. *What Do You Mean? Interpreting Image Classification with Crowdsourced Concept Extraction and Analysis*, 1937–1948. WWW '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383127.

Baron, R. M.; and Kenny, D. A. 1986. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6): 1173.

Brughmans, D.; Leyman, P.; and Martens, D. 2021. NICE: An Algorithm for Nearest Instance Counterfactual Explanations.

Buijsman, S. 2022. Defining Explanation and Explanatory Depth in XAI. *Minds and Machines*.

Chickering, D. M. 2002. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov): 507–554.

Cinelli, C.; Forney, A.; and Pearl, J. 2021. A crash course in good and bad controls. *Sociological Methods & Research*, 00491241221099552.

Dandl, S.; Molnar, C.; Binder, M.; and Bischl, B. 2020. Multi-Objective Counterfactual Explanations. In Bäck, T.; Preuss, M.; Deutz, A.; Wang, H.; Doerr, C.; Emmerich, M.; and Trautmann, H., eds., *Parallel Problem Solving from Nature – PPSN XVI*, 448–469. Cham: Springer International Publishing. ISBN 978-3-030-58112-1.

Daniusis, P.; Janzing, D.; Mooij, J.; Zscheischler, J.; Steudel, B.; Zhang, K.; and Schoelkopf, B. 2012. Inferring deterministic causal relations.

Draws, T.; Rieger, A.; Inel, O.; Gadiraju, U.; and Tintarev, N. 2021. A Checklist to Combat Cognitive Biases in Crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 9(1): 48–59.

Fonollosa, J. A. R. 2016. Conditional distribution variability measures for causality detection.

Freiesleben, T. 2021. The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples. *Minds and Machines*, 32(1): 77–109.

Freitas, A. A. 2014. Comprehensible Classification Models: A Position Paper. *SIGKDD Explor. Newsl.*, 15(1): 1–10.

Ghorbani, A.; Wexler, J.; Zou, J. Y.; and Kim, B. 2019. Towards Automatic Concept-based Explanations. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10.

Goyal, Y.; Wu, Z.; Ernst, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Counterfactual Visual Explanations.

Grimsley, C.; Mayfield, E.; and R.S. Bursten, J. 2020. Why Attention is Not Explanation: Surgical Intervention and Causal Reasoning about Neural Models. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 1780–1790. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Guidotti, R. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*.

Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; and Song, D. 2021. Natural Adversarial Examples. *CVPR*.

Hill, J.; and Stuart, E. A. 2015. Causal Inference: Overview. In Wright, J. D., ed., *International Encyclopedia of the Social and Behavioral Sciences (Second Edition)*, 255–260. Oxford: Elsevier, second edition edition. ISBN 978-0-08-097087-5.

Hoyer, P.; Janzing, D.; Mooij, J. M.; Peters, J.; and Schölkopf, B. 2008. Nonlinear causal discovery with additive noise models. In Koller, D.; Schuurmans, D.; Bengio, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Hyvärinen, A.; and Pajunen, P. 1999. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3): 429–439.

Iandola, F. N.; Han, S.; Moskewicz, M. W.; Ashraf, K.; Dally, W. J.; and Keutzer, K. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; and Sayres, R. 2017. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV).

Krishna, S.; Han, T.; Gu, A.; Pombra, J.; Jabbari, S.; Wu, S.; and Lakkaraju, H. 2022. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective.

Li, Y.; Huang, B.; Yang, H.; Hou, G.; Zhang, P.; and Duan, J. 2020. Efficient image structural similarity quality assessment method using image regularised feature. *IET Image Processing*, 14(16): 4401–4411.

Lundberg, S. M.; and Lee, S.-I. 2017. A Unified Approach to Interpreting Model Predictions. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38.

Mohseni, S.; Zarei, N.; and Ragan, E. D. 2018. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems.

Molnar, C. 2022. *Interpretable Machine Learning*. Lulu, 2 edition.

Paschos, G. 2001. Perceptually uniform color spaces for color texture analysis: An empirical evaluation. *Image Processing, IEEE Transactions on*, 10: 932 – 937.

Pearl, J. 1995. Causal Diagrams for Empirical Research. *Biometrika*, 82(4): 669–688.

Pearl, J. 2016. *Causal inference in statistics : a primer*. Chichester, West Sussex: Wiley. ISBN 9781119186847.

Pearl, J.; and Mackenzie, D. 2018. *The Book of Why: The New Science of Cause and Effect*. USA: Basic Books, Inc., 1st edition. ISBN 046509760X.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 1135–1144. New York, NY, USA: Association for Computing Machinery. ISBN 9781450342322.

Sackett, D. L. 1979. Bias in analytic research. *Journal of Chronic Diseases*, 32(1): 51–63.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2): 336–359.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2): 336–359.

Sharifi Noorian, S.; Qiu, S.; Gadiraju, U.; Yang, J.; and Bozzon, A. 2022. What Should You Know? A Human-In-the-Loop Approach to Unknown Unknowns Characterization in Image Recognition. In *Proceedings of the ACM Web Conference 2022*, WWW '22, 882–892. New York, NY, USA: Association for Computing Machinery. ISBN 9781450390965.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning Important Features Through Propagating Activation Differences. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 3145–3153. PMLR.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.

Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186. New York, NY, USA: Association for Computing Machinery. ISBN 9781450371100.

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. SmoothGrad: removing noise by adding noise.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT press.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic Attribution for Deep Networks. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 3319–3328. PMLR.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2015. Rethinking the Inception Architecture for Computer Vision.

Thagard, P. 1989. Explanatory coherence. *Behavioral and Brain Sciences*, 12(3): 435–467.

van der Waa, J.; Robeer, M.; van Diggelen, J.; Brinkhuis, M.; and Neerincx, M. 2018. Contrastive Explanations with Local Foil Trees.

VanderWeele, T. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR.

Weinberger, N. 2018. Faithfulness, Coordination and Causal Coincidences. *Erkenntnis*, 83(2): 113–133.

Westreich, D.; and Greenland, S. 2013. The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. *American Journal of Epidemiology*, 177(4): 292–298.

Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Xu, S.; Li, Y.; Liu, S.; Fu, Z.; Chen, X.; and Zhang, Y. 2020. Learning Post-Hoc Causal Explanations for Recommendation.

Yalcin, O.; Fan, X.; and Liu, S. 2021. Evaluating the Correctness of Explainable AI Algorithms for Classification.

Zhang, K.; Wang, Z.; Zhang, J.; and Schölkopf, B. 2015. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2): 1–22.

Zhang, L.; Sun, X.; Li, Y.; Zhang, Z.; and Feng, Y. 2019. A Noise-Sensitivity-Analysis-Based Test Prioritization Technique for Deep Neural Networks.

B

Fidelity Results

Model (Bias Type)	Class Name	Concepts (Effects)
Inception (Negative Set Bias)	Ants	plastic_box(0.68),notebook(0.68),leaf(0.64),wheel(0.57),bottle_cap(0.5)
	Mantis	dog(1.31),tree(0.65),mantis(0.61),storage_box(0.55),grasshopper(0.51)
	Bees	color_beige(0.76),camera(0.7),bicycle(0.67),bee(0.65),seal(0.63)
Inception (No Bias)	Tea Spoon	color_olive(0.75),color_olivedrab(0.75),spoon(0.74),photo(0.35),background(0.09)
	Fish Slice	paper(0.56),spatula(0.45),guitar(0.4),color_brown(0.39),color_gray(0.32)
	Dinner Cutlery	knife_handle(0.61),color_violet(0.46),knife(0.45),fork(0.42),color_bollywood(0.33)
Inception (Colour Bias)	Tea Spoon	teaspoon(0.62),color_yellow(0.5),color_green(0.41),spoon(0.39),color_bronze(0.2)
	Fish Slice	color_blue(0.55),spatula(0.49),guitar_keychain(0.32),color_gold(0.28),color_red(0.24)
	Dinner Cutlery	color_bronze(0.75),color_black(0.23),knife(0.22),butter_knife(0.2),color_golden(0.19)
Inception (Shape Bias)	Tea Spoon	color_khaki(0.19),color_grey(0.16),tablespoon(0.15),color_olive(0.14),color_bollywood(0.12)
	Fish Slice	color_steel(0.52),color_khaki(0.3),color_cyan(0.3),color_beige(0.3),butter_knife(0.21)
	Dinner Cutlery	color_darkgoldenrod(0.5),color_cornstalk(0.26),color_white(0.26),color_red(0.25),backgroud(0.17)
SqueezeNet (Negative Set Bias)	Ants	keyboard(1.0),grass(0.93),petri_dish(0.91),infested_trash_can(0.84),ant(0.82)
	Mantis	basket(0.97),photo(0.74),color_skin(0.73),house_windows(0.58),locust(0.56)
	Bees	drinking_fountain(1.35),water_bottle(0.67),backpack(0.66),glass_jar(0.6),color_orange(0.56)
SqueezeNet (No Bias)	Tea Spoon	color_plum(0.79),color_yellow(0.79),color_olive(0.78),wire(0.62),spoon(0.61)
	Fish Slice	color_bollywood(0.93),spatula(0.78),pasta_server(0.77),color_green(0.24),background(0.08)
	Dinner Cutlery	knife(0.6),fork(0.58),bulb(0.49),color_purple(0.33),color_azure(0.32)
SqueezeNet (Colour Bias)	Tea Spoon	spoon(0.69),color_yellow(0.62),color_olive(0.59),color_beige(0.53),teaspoon(0.4)
	Fish Slice	color_red(0.5),color_aqua(0.47),scraper(0.42),spatial(0.41),metal_drainer/spatula(0.38)
	Dinner Cutlery	color_palegoldenrod(0.71),butter_knife(0.56),knife(0.34),color_maroon(0.31),color_bronze(0.3)
SqueezeNet (Shape Bias)	Tea Spoon	spoon(0.66),teaspoon(0.64),color_red(0.56),color_olive(0.55),color_bollywood(0.4)
	Fish Slice	guitar(0.84),kitchen_spoon(0.52),spatula(0.5),utensil(0.47),color_goldenrod(0.27)
	Dinner Cutlery	fork(0.42),color_grey(0.32),color_black(0.26),color_goldenrod(0.26),butter_knife(0.21)

Table B.1: Fidelity results for template based causal discovery with O-C data selection strategy

Inception (Negative Set Bias)	Ants	plastic_box(0.62),leaf(0.62),notebook(0.62),wheel(0.54),bottle_cap(0.47)
	Mantis	dog(1.27),tree(0.61),mantis(0.55),storage_box(0.52),clock(0.49)
	Bees	color_beige(0.74),camera(0.71),bicycle(0.68),bee(0.65),seal(0.61)
Inception (No Bias)	Tea Spoon	spoon(0.73),color_olive(0.71),color_olivedrab(0.7),photo(0.39),shape_cylinder(0.2)
	Fish Slice	paper(0.57),guitar(0.5),spatula(0.48),color_brown(0.41),shape_square(0.3)
	Dinner Cutlery	knife_handle(0.58),knife(0.44),fork(0.37),color_bollywood(0.31),color_white(0.18)
Inception (Colour Bias)	Tea Spoon	teaspoon(0.62),color_green(0.46),color_yellow(0.43),spoon(0.39),color_bronze(0.22)
	Fish Slice	spatula(0.48),color_blue(0.47),guitar_keychain(0.4),color_gold(0.3),fish_knife(0.22)
	Dinner Cutlery	color_bronze(0.74),color_black(0.32),knife(0.22),butter_knife(0.2),color_brown(0.18)
Inception (Shape Bias)	Tea Spoon	color_grey(0.25),color_khaki(0.2),shape_rectangular(0.17),tablespoon(0.15),color_olive(0.15)
	Fish Slice	color_steel(0.53),color_khaki(0.3),shape_square(0.29),butter_knife(0.22),color_beige(0.18)
	Dinner Cutlery	color_darkgoldenrod(0.51),color_red(0.27),color_white(0.25),color_blue(0.18),knife(0.16)
SqueezeNet (Negative Set Bias)	Ants	grass(0.89),petri_dish(0.84),infested_trash_can(0.76),keyboard(0.72),ant(0.72)
	Mantis	basket(1.0),photo(0.75),color_skin(0.7),ant_statue(0.62),prunes(0.62)
	Bees	water_bottle(0.68),backpack(0.68),drinking_fountain(0.65),glass_jar(0.64),bee(0.6)
SqueezeNet (No Bias)	Tea Spoon	color_plum(0.75),color_olive(0.75),wire(0.62),string(0.57),teaspoon(0.57)
	Fish Slice	color_bollywood(0.97),spatula(0.78),pasta_server(0.68),shape_rectangle(0.31),color_red(0.27)
	Dinner Cutlery	knife(0.6),fork(0.58),color_gold(0.36),bulb(0.36),color_white(0.33)
SqueezeNet (Colour Bias)	Tea Spoon	color_beige(1.19),shape_long(0.81),spoon(0.77),color_yellow(0.67),teaspoon(0.48)
	Fish Slice	spatial(0.51),color_aqua(0.45),color_red(0.42),scraper(0.4),color_white(0.37)
	Dinner Cutlery	butter_knife(0.6),color_palegoldenrod(0.4),shape_square(0.37),knife(0.35),color_bronze(0.32)
SqueezeNet (Shape Bias)	Tea Spoon	teaspoon(0.68),spoon(0.67),color_olive(0.56),color_red(0.34),color_bollywood(0.33)
	Fish Slice	guitar(0.84),kitchen_spoon(0.48),spatula(0.48),utensil(0.44),color_yellow(0.27)
	Dinner Cutlery	fork(0.42),color_grey(0.37),color_goldenrod(0.32),shape_curve(0.29),color_black(0.29)

Table B.2: Fidelity results for template-based causal discovery with O-C-S data selection strategy

Inception (Negative Set Bias)	Ants	middle_of_notebook(0.71),plastic_box(0.71),bottom_of_leaf(0.65),middle_of_wheel(0.61),outercircle_of_wheel(0.59)
	Mantis	legs_of_dog(1.35),top-left_of_tree(0.7),head_of_grasshopper(0.66),mantis(0.64),tree(0.62)
	Bees	color_beige(0.75),camera(0.7),middle_of_bicycle(0.67),bee(0.67),seal(0.67)
Inception (No Bias)	Tea Spoon	top-left_of_spoon(0.92),head_of_spoon(0.7),color_olive(0.69),color_olivedrab(0.69),handle_of_spoon(0.69)
	Fish Slice	shade_of_spatula(0.7),side_of_spatula(0.63),paper(0.57),head_of_spatula(0.52),spatula(0.47)
	Dinner Cutlery	knife_handle(0.61),bottom-left_of_background(0.53),knife(0.5),fork(0.43),tail_of_fork(0.42)
Inception (Colour Bias)	Tea Spoon	head_of_teaspoon(0.59),color_yellow(0.5),tail_of_teaspoon(0.48),light_reflections_of_spoon(0.42),handle_of_spoon(0.4)
	Fish Slice	bottom_of_spatula(0.56),middle_of_the_handle_of_spatula(0.56),handle_of_spatula(0.54),color_blue(0.51),spatula(0.49)
	Dinner Cutlery	color_bronze(0.75),middle_of_background(0.55),light_reflections_of_background(0.31),texture_of_background(0.31),color_black(0.24)
Inception (Shape Bias)	Tea Spoon	border_of_spoon(0.23),head_of_fork(0.23),handle_of_fork(0.23),bottom_of_background(0.17),color_olive(0.14)
	Fish Slice	bottom_of_fork(0.58),color_steel(0.44),bottom_of_background(0.35),butter_knife(0.27),color_cyan(0.25)
	Dinner Cutlery	middle_of_background(0.26),color_white(0.26),fork(0.25),color_golden(0.23),top_of_spoon(0.23)
SqueezeNet (Negative Set Bias)	Ants	copyright_of_petri_dish(0.89),side_of_leaf(0.89),bottom_right_of_grass(0.86),color_plum(0.78),bottom_of_petri_dish(0.74)
	Mantis	inside_of_basket(1.04),string_of_tree(1.03),color_skin(0.74),background_of_photo(0.68),middle_of_insect(0.64)
	Bees	bottom_of_glass_jar(0.75),water_bottle(0.73),backpack(0.72),top_of_drinking_fountain(0.69),top_of_can(0.66)
SqueezeNet (No Bias)	Tea Spoon	color_plum(0.8),color_yellow(0.8),color_olive(0.8),spoon_light_reflection_of_spoon(0.68),wire(0.66)
	Fish Slice	color_bollywood(0.93),handle_of_spatula(0.76),spatula(0.76),pasta_server(0.75),head_of_spatula(0.74)
	Dinner Cutlery	middle_of_fork(0.75),edge_of_knife(0.73),knife(0.52),fork(0.5),top-left_of_knife(0.49)
SqueezeNet (Colour Bias)	Tea Spoon	spoon(0.67),head_of_fork(0.63),color_olive(0.59),handle_of_knife(0.58),teaspoon(0.5)
	Fish Slice	light_reflections_of_fork(0.54),bottom_of_scraper(0.47),color_red(0.46),color_green(0.44),metal_drainer/spatula(0.43)
	Dinner Cutlery	color_palegoldenrod(0.8),background_of_knife(0.61),butter_knife(0.52),bottom_of_knife(0.42),head_of_knife(0.42)
SqueezeNet (Shape Bias)	Tea Spoon	head_of_spoon(0.86),tail_of_spoon(0.79),color_red(0.74),color_bollywood(0.74),teaspoon(0.57)
	Fish Slice	top_of_spatula(0.87),guitar(0.87),shadow_of_spatula(0.86),handle_of_spatula(0.58),head_of_spatula(0.56)
	Dinner Cutlery	top_of_background(0.73),color_grey(0.59),head_of_fork(0.48),fork(0.41),color_black(0.36)

Table B.3: Fidelity results for template-based causal discovery with PO-C data selection strategy

Inception (Negative Set Bias)	Ants	middle_of_notebook(0.66),plastic_box(0.66),bottom_of_leaf(0.64),middle_of_wheel(0.56),bottle_cap(0.54)
	Mantis	legs_of_dog(1.24),top-left_of_tree(0.59),head_of_grasshopper(0.55),tree(0.55),mantis(0.52)
	Bees	color_beige(0.72),camera(0.72),bee(0.69),middle_of_bicycle(0.69),seal(0.66)
Inception (No Bias)	Tea Spoon	top-left_of_spoon(0.83),head_of_spoon(0.68),color_olive(0.67),color_olivedrab(0.67),handle_of_spoon(0.67)
	Fish Slice	shade_of_spatula(0.85),side_of_spatula(0.68),spatula(0.66),guitar(0.62),color_brown(0.6)
	Dinner Cutlery	knife_handle(0.64),bottom-left_of_background(0.55),knife(0.5),fork(0.44),tail_of_fork(0.41)
Inception (Colour Bias)	Tea Spoon	handle_of_spoon(0.64),head_of_teaspoon(0.6),tail_of_teaspoon(0.5),color_yellow(0.45),color_green(0.44)
	Fish Slice	middle_of_the_handle_of_spatula(0.74),bottom_of_spatula(0.73),handle_of_spatula(0.56),spatula(0.51),guitar_keychain(0.5)
	Dinner Cutlery	color_bronze(0.69),color_black(0.53),middle_of_background(0.42),knife(0.26),butter_knife(0.25)
Inception (Shape Bias)	Tea Spoon	border_of_spoon(0.27),handle_of_fork(0.2),head_of_fork(0.2),color_olive(0.16),bottom_of_background(0.15)
	Fish Slice	bottom_of_fork(0.66),bottom_of_background(0.43),butter_knife(0.38),color_steel(0.37),tablespoon(0.26)
	Dinner Cutlery	top_of_spoon(0.39),color_golden(0.25),color_red(0.23),color_goldenrod(0.22),middle_of_background(0.22)
SqueezeNet (Negative Set Bias)	Ants	copyright_of_petri_dish(0.97),bottom_right_of_grass(0.96),side_of_leaf(0.88),bottom_of_petri_dish(0.76),color_plum(0.75)
	Mantis	string_of_tree(1.04),inside_of_basket(1.03),color_skin(0.75),background_of_photo(0.68),middle_of_insect(0.62)
	Bees	bottom_of_glass_jar(0.72),water_bottle(0.7),backpack(0.69),top_of_drinking_fountain(0.66),bee(0.63)
SqueezeNet (No Bias)	Tea Spoon	color_plum(0.81),color_olive(0.8),spoon(0.64),wire(0.64),teaspoon(0.64)
	Fish Slice	color_bollywood(0.94),handle_of_spatula(0.75),top_of_spatula(0.74),spatula(0.74),head_of_spatula(0.71)
	Dinner Cutlery	middle_of_fork(0.77),edge_of_knife(0.75),top-left_of_knife(0.71),top_of_fork(0.67),knife(0.54)
SqueezeNet (Colour Bias)	Tea Spoon	shape_long(1.12),spoon(0.71),handle_of_knife(0.64),head_of_fork(0.64),color_yellow(0.63)
	Fish Slice	light_reflections_of_fork(0.46),color_red(0.44),bottom-right-corner_of_spatial(0.4),top_of_scraper(0.4),tail_of_fork(0.39)
	Dinner Cutlery	background_of_knife(0.59),butter_knife(0.55),head_of_knife(0.49),bottom_of_knife(0.49),color_palegoldenrod(0.4)
SqueezeNet (Shape Bias)	Tea Spoon	color_red(0.95),color_bollywood(0.95),head_of_spoon(0.87),tail_of_spoon(0.83),teaspoon(0.67)
	Fish Slice	top_of_spatula(0.78),guitar(0.78),shadow_of_spatula(0.77),head_of_spatula(0.54),bottom_of_spatula(0.51)
	Dinner Cutlery	top_of_background(0.93),color_grey(0.65),head_of_fork(0.51),color_black(0.41),color_goldenrod(0.4)

Table B.4: Fidelity results for template-based causal discovery with PO-C-S data selection strategy

Inception (Negative Set Bias)	Ants	color_chocolate(0.66),wheel(0.63),bottle_cap(0.52),branding_paper(0.52),trash_bag(0.51)
	Mantis	dog(0.87),tree(0.16),color_gold(0.12),mantis(0.11),storage_box(0.06)
	Bees	bicycle(0.64),bee(0.5),color_pink(0.38),camera(0.35),color_clear(0.23)
Inception (No Bias)	Tea Spoon	spoon(0.64),color_olive(0.52),color_olivedrab(0.52),photo(0.22),color_green(0.07)
	Fish Slice	color_brown(0.44),spatula(0.43),color_black(0.43),paper(0.32),color_gray(0.23)
	Dinner Cutlery	fork(0.36),color_silver(0.31),knife(0.3),knife_handle(0.23),color_violet(0.19)
Inception (Colour Bias)	Tea Spoon	teaspoon(0.73),spoon(0.56),fork(0.35),color_green(0.3),knife(0.29)
	Fish Slice	color_blue(0.54),color_red(0.2),color_grey(0.18),guitar_keychain(0.16),color_gold(0.16)
	Dinner Cutlery	color_black(0.38),color_bronze(0.32),background(0.23),color_silver(0.22),knife(0.22)
Inception (Shape Bias)	Tea Spoon	color_grey(0.22),color_khaki(0.09),color_olive(0.08),tablespoon(0.08),fork(0.06)
	Fish Slice	color_steel(0.3),color_beige(0.28),color_khaki(0.22),color_cyan(0.22),color_darkviolet(0.06)
	Dinner Cutlery	spatula(0.33),color_green(0.3),knife(0.25),background(0.22),color_darkgoldenrod(0.22)
SqueezeNet (Negative Set Bias)	Ants	ant(0.58),color_gray(0.54),color_maroon(0.54),color_plum(0.22),color_skin(0.01)
	Mantis	basket(0.52),color_skin(0.5),color_yellow(0.31),ant_statue(0.26),cups(0.25)
	Bees	color_transplant(0.84),fingers(0.83),can(0.79),bee(0.78),bicycle_bidon(0.77)
SqueezeNet (No Bias)	Tea Spoon	spoon(0.54),teaspoon(0.37),wire(0.36),color_yellow(0.36),color_green(0.35)
	Fish Slice	spatula(0.72),color_bollywood(0.49),pasta_server(0.48),color_green(0.21),color_red(0.07)
	Dinner Cutlery	color_purple(0.76),knife(0.7),color_black(0.66),background(0.66),fork(0.63)
SqueezeNet (Colour Bias)	Tea Spoon	color_yellow(0.53),color_olive(0.43),color_grey(0.38),color_beige(0.36),color_gold(0.32)
	Fish Slice	scraper(0.25),color_red(0.22),color_green(0.2),utensil(0.18),spatial(0.18)
	Dinner Cutlery	color_palegoldenrod(0.75),color_brown(0.47),color_darkgray(0.3),knife(0.23),color_bronze(0.23)
SqueezeNet (Shape Bias)	Tea Spoon	teaspoon(0.64),spoon(0.63),color_green(0.34),color_red(0.33),knife(0.27)
	Fish Slice	spatula(0.66),guitar(0.51),kitchen_spoon(0.48),utensil(0.45),color_silver(0.22)
	Dinner Cutlery	color_grey(0.27),color_black(0.26),fork(0.19),color_goldenrod(0.16),butter_knife(0.13)

Table B.5: Fidelity results for CDS causal discovery with O-C data selection strategy

Inception (Negative Set Bias)	Ants	wheel(0.28),shape_oval(0.15),shape_irregular(0.14),color_chocolate(0.14),color_clear(0.12)
	Mantis	dog(0.71),color_grey(0.63),shape_rectangular(0.37),color_white(0.1),color_gold(0.1)
	Bees	camera(0.69),bee(0.49),shape_triangle(0.48),finger(0.47),furniture(0.47)
Inception (No Bias)	Tea Spoon	spoon(0.6),color_olivedrab(0.44),color_olive(0.44),photo(0.22),shape_rectangle(0.16)
	Fish Slice	color_brown(0.52),spatula(0.5),color_black(0.34),paper(0.3),guitar(0.27)
	Dinner Cutlery	color_silver(0.35),knife(0.3),fork(0.28),knife_handle(0.21),color_bollywood(0.17)
Inception (Colour Bias)	Tea Spoon	teaspoon(0.64),fork(0.59),spoon(0.58),color_green(0.56),knife(0.48)
	Fish Slice	color_grey(0.53),color_blue(0.45),color_gold(0.16),guitar_keychain(0.16),color_red(0.15)
	Dinner Cutlery	color_black(0.41),color_bronze(0.32),color_silver(0.2),background(0.19),color_brown(0.09)
Inception (Shape Bias)	Tea Spoon	color_grey(0.23),shape_rectangular(0.14),tablespoon(0.11),color_khaki(0.1),color_olive(0.08)
	Fish Slice	color_steel(0.3),color_khaki(0.21),shape_square(0.21),color_beige(0.21),fork(0.12)
	Dinner Cutlery	color_darkgoldenrod(0.22),knife(0.17),color_blue(0.17),color_red(0.14),backgroud(0.11)
SqueezeNet (Negative Set Bias)	Ants	shape_splotches(0.66),grass(0.62),gravel_road(0.62),ant(0.62),shape_round(0.5)
	Mantis	basket(0.5),color_skin(0.5),shape_crescent(0.5),ant_statue(0.49),shape_circus(0.45)
	Bees	color_transplant(0.84),shape_rectangle(0.84),backpack(0.83),fingers(0.83),bee(0.79)
SqueezeNet (No Bias)	Tea Spoon	spoon(0.38),wire(0.37),teaspoon(0.36),string(0.33),color_plum(0.33)
	Fish Slice	color_bollywood(0.82),spatula(0.79),shape_rectangle(0.35),pasta_server(0.35),color_green(0.33)
	Dinner Cutlery	knife(0.84),color_purple(0.83),color_black(0.76),fork(0.62),background(0.5)
SqueezeNet (Colour Bias)	Tea Spoon	color_beige(0.89),shape_long(0.72),spoon(0.66),color_olive(0.43),shape_rectangular(0.4)
	Fish Slice	color_blue(0.42),scraper(0.28),color_aqua(0.27),spatial(0.26),shape_line(0.21)
	Dinner Cutlery	color_palegoldenrod(0.37),shape_square(0.37),knife(0.33),color_darkgray(0.31),color_black(0.3)
SqueezeNet (Shape Bias)	Tea Spoon	spoon(0.55),color_green(0.33),teaspoon(0.3),color_red(0.29),color_bollywood(0.23)
	Fish Slice	spatula(0.63),kitchen_spoon(0.61),utensil(0.49),guitar(0.48),color_yellow(0.37)
	Dinner Cutlery	knife(0.33),color_grey(0.3),color_black(0.25),fork(0.24),background(0.23)

Table B.6: Fidelity results for CDS causal discovery with O-C-S data selection strategy

Inception (Negative Set Bias)	Ants	color_chocolate(0.65),bottle_cap(0.61),branding_paper(0.61),color_red(0.5),color_yellow(0.47)
	Mantis	legs_of_dog(0.9),top-left_of_tree(0.22),head_of_grasshopper(0.18),mantis(0.16),tree(0.14)
	Bees	bee(0.4),color_pink(0.36),camera(0.35),color_clear(0.34),bottom_of_furniture(0.25)
Inception (No Bias)	Tea Spoon	top-left_of_spoon(0.95),head_of_spoon(0.69),handle_of_spoon(0.69),spoon(0.45),color_olive(0.44)
	Fish Slice	paper(0.69),spatula(0.66),head_of_spatula(0.45),side_of_spatula(0.44),shade_of_spatula(0.44)
	Dinner Cutlery	fork(0.45),knife(0.4),spikes_of_fork(0.33),tail_of_fork(0.33),head_of_knife(0.29)
Inception (Colour Bias)	Tea Spoon	head_of_teaspoon(0.64),light_reflections_of_spoon(0.52),spoon(0.51),head_of_spoon(0.44),handle_of_spoon(0.39)
	Fish Slice	spatula(0.38),handle_of_spatula(0.37),color_blue(0.32),bottom_of_spatula(0.3),middle_of_the_handle_of_spatula(0.3)
	Dinner Cutlery	color_black(0.64),color_bronze(0.32),color_silver(0.17),knife(0.15),light_reflections_of_background(0.12)
Inception (Shape Bias)	Tea Spoon	handle_of_fork(0.14),head_of_fork(0.14),border_of_spoon(0.12),bottom_of_background(0.1),tablespoon(0.08)
	Fish Slice	bottom_of_fork(0.39),color_steel(0.3),bottom_of_background(0.21),color_beige(0.19),color_cyan(0.18)
	Dinner Cutlery	top_of_spatula(0.27),top_of_spoon(0.22),knife(0.2),color_blue(0.19),shadow(0.18)
SqueezeNet (Negative Set Bias)	Ants	bottom_of_petri_dish(0.45),copyright_of_petri_dish(0.41),bottom_right_of_grass(0.39),ant(0.39),color_maroon(0.28)
	Mantis	color_skin(0.5),basket(0.5),color_plum(0.46),ant_statue(0.33),cups(0.31)
	Bees	color_transplant(0.92),color_red(0.9),color_orange(0.76),color_grey(0.57),bottom_of_glass_jar(0.52)
SqueezeNet (No Bias)	Tea Spoon	spoon(0.53),wire(0.43),light_reflection_of_spoon(0.42),teaspoon(0.4),spoon_light_reflection_of_spoon(0.38)
	Fish Slice	handle_of_spatula(0.66),spatula(0.64),head_of_spatula(0.55),color_bollywood(0.46),pasta_server(0.4)
	Dinner Cutlery	color_black(0.65),background(0.64),color_purple(0.53),color_steel(0.28),color_azure(0.26)
SqueezeNet (Colour Bias)	Tea Spoon	color_yellow(0.43),color_grey(0.37),color_olive(0.36),spoon(0.28),color_silver(0.15)
	Fish Slice	spatula(0.49),bottom_of_scraper(0.42),metal_drainer/spatula(0.41),light_reflections_of_fork(0.39),bottom_of_utensil(0.34)
	Dinner Cutlery	color_palegoldenrod(0.45),color_black(0.27),color_maroon(0.24),color_bronze(0.21),background_of_fork(0.2)
SqueezeNet (Shape Bias)	Tea Spoon	tail_of_spoon(0.5),spoon(0.39),color_red(0.39),teaspoon(0.39),color_bollywood(0.38)
	Fish Slice	top_of_spatula(0.49),guitar(0.47),shadow_of_spatula(0.47),head_of_spatula(0.32),bottom_of_spatula(0.32)
	Dinner Cutlery	color_blue(0.34),color_goldenrod(0.17),top_of_background(0.16),color_grey(0.15),color_silver(0.1)

Table B.7: Fidelity results for CDS causal discovery with PO-C data selection strategy

Inception (Negative Set Bias)	Ants	branding_paper(0.39),bottle_cap(0.39),shape_oval(0.32),bottom-right-corner_of_pump(0.29),shape_circle(0.27)
	Mantis	legs_of_dog(0.7),color_grey(0.38),shape_rectangular(0.15),color_gold(0.08),window(0.05)
	Bees	shape_tubular(0.72),shape_triangle(0.47),bottom_of_furniture(0.46),color_pink(0.46),bee(0.45)
Inception (No Bias)	Tea Spoon	spoon(0.65),head_of_spoon(0.62),handle_of_spoon(0.52),color_olive(0.37),color_olivedrab(0.37)
	Fish Slice	color_brown(0.62),spatula(0.52),shade_of_spatula(0.47),color_black(0.47),guitar(0.36)
	Dinner Cutlery	color_silver(0.55),fork(0.42),knife(0.32),knife_handle(0.23),background_of_photo(0.22)
Inception (Colour Bias)	Tea Spoon	light_reflections_of_spoon(0.56),spoon(0.52),fork(0.44),handle_of_spoon(0.42),color_green(0.4)
	Fish Slice	color_blue(0.4),color_grey(0.35),middle_of_the_handle_of_spatula(0.32),bottom_of_spatula(0.28),handle_of_spatula(0.2)
	Dinner Cutlery	color_black(0.69),color_bronze(0.32),color_silver(0.25),color_green(0.15),shape_cylindrical(0.11)
Inception (Shape Bias)	Tea Spoon	border_of_spoon(0.11),bottom_of_background(0.1),tablespoon(0.08),shape_rectangular(0.08),color_grey(0.06)
	Fish Slice	color_steel(0.3),bottom_of_fork(0.26),bottom_of_background(0.22),color_beige(0.2),shape_square(0.17)
	Dinner Cutlery	tail_of_spatula(0.36),copyright_of_spoon(0.28),spatula(0.28),top_of_spoon(0.28),shadow(0.28)
SqueezeNet (Negative Set Bias)	Ants	shape_splotches(0.77),bottom_of_petri_dish(0.75),ant(0.74),bottom_right_of_grass(0.73),gravel_road(0.73)
	Mantis	color_skin(0.5),shape_crescent(0.5),ant_statue(0.49),shape_circus(0.43),basket(0.39)
	Bees	color_transplant(0.8),shape_rectangle(0.8),fingers(0.79),shape_circle(0.78),bee(0.74)
SqueezeNet (No Bias)	Tea Spoon	spoon(0.52),light_reflection_of_spoon(0.41),teaspoon(0.39),wire(0.37),color_plum(0.36)
	Fish Slice	spatula(0.65),handle_of_spatula(0.56),head_of_spatula(0.48),color_bollywood(0.46),top_of_spatula(0.4)
	Dinner Cutlery	color_purple(0.73),color_black(0.48),background(0.34),color_gold(0.29),shape_rectangle(0.24)
SqueezeNet (Colour Bias)	Tea Spoon	shape_long(0.88),color_yellow(0.35),color_grey(0.33),spoon(0.31),shape_rectangular(0.27)
	Fish Slice	spatula(0.39),bottom_of_scraper(0.37),metal_drainer/spatula(0.36),top_of_scraper(0.33),fork(0.31)
	Dinner Cutlery	shape_square(0.36),color_palegoldenrod(0.3),color_darkgray(0.26),color_black(0.26),knife(0.24)
SqueezeNet (Shape Bias)	Tea Spoon	color_bollywood(0.55),color_red(0.54),teaspoon(0.39),tail_of_spoon(0.25),color_green(0.2)
	Fish Slice	top_of_spatula(0.45),guitar(0.45),shadow_of_spatula(0.44),head_of_spatula(0.32),bottom_of_spatula(0.31)
	Dinner Cutlery	shape_curve(0.41),spoon(0.39),fork(0.24),top_of_background(0.24),shape_semicircle(0.24)

Table B.8: Fidelity results for CDS causal discovery with PO-C-S data selection strategy

Bibliography

Abrevaya, J., Hsu, Y.-C., & Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics*, *33*(4), 485–505.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, *10*(7), 1–46. https://doi.org/10.1371/journal.pone.0130140

Balayn, A., Soilis, P., Lofi, C., Yang, J., & Bozzon, A. (2021). What do you mean? interpreting image classification with crowdsourced concept extraction and analysis. *Proceedings of the web conference 2021* (pp. 1937–1948). Association for Computing Machinery. https://doi.org/10.1145/3442381.3450069

Balayn, A., Lofi, C., & Houben, G. (2021). Managing bias and unfairness in data for decision support: A survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *The VLDB Journal*, *30*(5), 739–768. https://doi.org/10.1007/s00778-021-00671-8

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, *51*(6), 1173.

Barros, D. B. (2013). Negative causation in causal and mechanistic explanation. *Synthese*, *190*(3), 449–469. https://doi.org/10.1007/s11229-011-0040-4

Bergstrom, C., & West, J. (2016). Case study — criminal machine learning [(Accessed on 10/31/2021)].

Bezdek, J. C., Keller, J., Krisnapuram, R., & Pal, N. R. (1999). Image processing and computer vision. *The handbooks of fuzzy sets series* (pp. 547–678). The Handbooks of Fuzzy Sets Series. https://doi.org/10.1007/0-387-24579-0_5

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. https://doi.org/10.48550/ARXIV.2005.14165

Brughmans, D., Leyman, P., & Martens, D. (2021). Nice: An algorithm for nearest instance counterfactual explanations. https://doi.org/10.48550/ARXIV.2104.07411

Buijsman, S. (2022). Defining explanation and explanatory depth in xai. *Minds and Machines*. https://doi.org/10.1007/s11023-022-09607-9

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, *3*(Nov), 507–554.

Cinelli, C., Forney, A., & Pearl, J. (2021). A crash course in good and bad controls. *Sociological Methods & Research*, 00491241221099552.

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020). Multi-objective counterfactual explanations. In T. Bäck, M. Preuss, A. Deutz, H. Wang, C. Doerr, M. Emmerich, & H. Trautmann (Eds.), *Parallel problem solving from nature – ppsn xvi* (pp. 448–469). Springer International Publishing.

Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., & Schoelkopf, B. (2012). Inferring deterministic causal relations. https://doi.org/10.48550/ARXIV.1203.3475

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. https://doi.org/10.1109/cvpr.2009.5206848

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. https://doi.org/10.48550/ARXIV.1702.08608

Draws, T., Rieger, A., Inel, O., Gadiraju, U., & Tintarev, N. (2021). A checklist to combat cognitive biases in crowdsourcing. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *9*(1), 48–59. https://ojs.aaai.org/index.php/HCOMP/article/view/18939

Everingham, M., Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, *88*(2), 303–338. https://doi.org/10.1007/s11263-009-0275-4

Fonollosa, J. A. R. (2016). Conditional distribution variability measures for causality detection. https://doi.org/10.48550/ARXIV.1601.06680

Freiesleben, T. (2021). The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines*, *32*(1), 77–109. https://doi.org/10.1007/s11023-021-09580-9

Freitas, A. A. (2014). Comprehensible classification models: A position paper. *SIGKDD Explor. Newsl.*, *15*(1), 1–10. https://doi.org/10.1145/2594473.2594475

Ghorbani, A., Wexler, J., Zou, J. Y., & Kim, B. (2019). Towards automatic concept-based explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00524

Goren, A., Vano-Galvan, S., Wambier, C. G., McCoy, J., Gomez-Zubiaur, A., Moreno-Arrones, O. M., Shapiro, J., Sinclair, R. D., Gold, M. H., Kovacevic, M., et al. (2020). A preliminary observation: Male pattern hair loss among hospitalized covid-19 patients in spain-a potential clue to the role of androgens in covid-19 severity. *J Cosmet Dermatol*, *19*(7), 1545–1547.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019). Counterfactual visual explanations. https://doi.org/10.48550/ARXIV.1904.07451

Greene, T. (2020). 2010 – 2019: The rise of deep learning [(Accessed on 08/23/2022)].

Griffith, G. J., Morris, T. T., Tudball, M. J., Herbert, A., Mancano, G., Pike, L., Sharp, G. C., Sterne, J., Palmer, T. M., Davey Smith, G., & et al. (2020). Collider bias undermines our understanding of covid-19 disease risk and severity. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19478-2

Grimsley, C., Mayfield, E., & R.S. Bursten, J. (2020). Why attention is not explanation: Surgical intervention and causal reasoning about neural models. *Proceedings of the 12th Language Resources and Evaluation Conference*, 1780–1790. https://aclanthology.org/2020.lrec-1.220

Guidotti, R. (2022). Counterfactual explanations and how to find them: Literature review and benchmarking. *Data Mining and Knowledge Discovery*. https://doi.org/10.1007/s10618-022-00831-6

Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials - the gold standard for effectiveness research. *BJOG: An International Journal of Obstetrics & Gynaecology*, *125*(13), 1716–1716. https://doi.org/10.1111/1471-0528.15199

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). Natural adversarial examples. *CVPR*.

Hill, J., & Stuart, E. A. (2015). Causal inference: Overview. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences (second edition)* (Second Edition, pp. 255–260). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-08-097086-8.42095-7

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems*. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2008/file/f7664060cc52bc6f3d620bcedc94a4b6-Paper.pdf

Huang, T. (1996). Computer vision: Evolution and promise.

Hudgens, M. G., & Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, *103*(482), 832–842. https://doi.org/10.1198/016214508000000292

Hyvärinen, A., & Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, *12*(3), 429–439.

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. https://doi.org/10.48550/ARXIV.1602.07360

Jo, J., & Bengio, Y. (2017). Measuring the tendency of cnns to learn surface statistical regularities. https://doi.org/10.48550/ARXIV.1711.11561

Kalainathan, D., & Goudet, O. (2019). Causal discovery toolbox: Uncover causal relationships in python. https://doi.org/10.48550/ARXIV.1903.02278

Kaushik, D., Hovy, E., & Lipton, Z. C. (2019). Learning the difference that makes a difference with counterfactually-augmented data. https://doi.org/10.48550/ARXIV.1909.12434

Kaushik, D., Setlur, A., Hovy, E., & Lipton, Z. C. (2020). Explaining the efficacy of counterfactually augmented data. https://doi.org/10.48550/ARXIV.2010.02114

Kidd, C., & Benjamin. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449–460. https://doi.org/10.1016/j.neuron.2015.09.010

Kim, B., Koyejo, O., & Khanna, R. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *NIPS*.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2017). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). https://doi.org/10.48550/ARXIV.1711.11279

Lee, N. T., Resnick, P., & Barton, G. (2019). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. *Brookings Institute: Washington, DC, USA*.

Li, Y., Huang, B., Yang, H., Hou, G., Zhang, P., & Duan, J. (2020). Efficient image structural similarity quality assessment method using image regularised feature. *IET Image Processing*, *14*(16), 4401–4411. https://doi.org/10.1049/iet-ipr.2019.1570

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2014). Microsoft coco: Common objects in context. https://doi.org/10.48550/ARXIV.1405.0312

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplement*, *27*, 247–266. https://doi.org/10.1017/S1358246100005130

Maccoun, R. J. (1998). Biases in the interpretation and use of research results. *Annual Review of Psychology*, *49*(1), 259–287. https://doi.org/10.1146/annurev.psych.49.1.259

McGregor, S. (2021). Preventing repeated real world ai failures by cataloging incidents: The ai incident database. *AAAI*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *CoRR*, *abs/1908.09635*. http://arxiv.org/abs/1908.09635

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

Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A multidisciplinary survey and framework for design and evaluation of explainable ai systems. https://doi.org/10.48550/ARXIV.1811.11839

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). Lulu. https://christophm.github.io/interpretable-ml-book

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

Neal, B. (2020). Introduction to causal inference.

O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. https://doi.org/10.48550/ARXIV.1511.08458

Paschos, G. (2001). Perceptually uniform color spaces for color texture analysis: An empirical evaluation. *Image Processing, IEEE Transactions on*, *10*, 932–937. https://doi.org/10.1109/83.923289

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2016). *Causal inference in statistics : A primer*. Wiley.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect* (1st). Basic Books, Inc.

Ranney, M. A., & Thagard, P. (1988). Explanatory coherence and belief revision in naive physics.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, *100*(469), 322–331. https://doi.org/10.1198/016214504000001880

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Sharifi Noorian, S., Qiu, S., Gadiraju, U., Yang, J., & Bozzon, A. (2022). What should you know? a human-in-the-loop approach to unknown unknowns characterization in image recognition. *Proceedings of the ACM Web Conference 2022*, 882–892. https://doi.org/10.1145/3485447.3512040

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (pp. 3145–3153). PMLR. https://proceedings.mlr.press/v70/shrikumar17a.html

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. https://doi.org/10.48550/ARXIV.1312.6034

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *13*(2), 238–241. Retrieved August 6, 2022, from http://www.jstor.org/stable/2984065

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: Removing noise by adding noise. https://doi.org/10.48550/ARXIV.1706.03825

Spirtes, P., Glymour, C. N., Scheines, R., & Heckerman, D. (2000). *Causation, prediction, and search*. MIT press.

Sprenger, J., & Weinberger, N. (2021). Simpson's Paradox. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the inception architecture for computer vision. https://doi.org/10.48550/ARXIV.1512.00567

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*(3), 435–467. https://doi.org/10.1017/S0140525X00057046

Thiele, T., Sommer, T., Schröder, S., Richert, A., & Jeschke, S. (2016). Human-in-the-loop processes as enabler for data analytics in digitalized organizations. *Mensch & Computer Workshopband*.

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *CVPR 2011*, 1521–1528. https://doi.org/10.1109/CVPR.2011.5995347

van der Waa, J., Robeer, M., van Diggelen, J., Brinkhuis, M., & Neerincx, M. (2018). Contrastive explanations with local foil trees. https://doi.org/10.48550/ARXIV.1806.07470

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. https://doi.org/10.48550/ARXIV.1711.00399

Weinberger, N. (2018). Faithfulness, coordination and causal coincidences. *Erkenntnis*, *83*(2), 113–133. https://doi.org/10.1007/s10670-017-9882-6

Westreich, D., & Greenland, S. (2013). The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients. *American Journal of Epidemiology*, *177*(4), 292–298. https://doi.org/10.1093/aje/kws412

Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C. E., Curtis, H. J., Mehrkar, A., Evans, D., Inglesby, P., & et al. (2020). Factors associated with covid-19-related death using opensafely. *Nature*, *584*(7821), 430–436. https://doi.org/10.1038/s41586-020-2521-4

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Wu, X., & Zhang, X. (2016). Responses to critiques on machine learning of criminality perceptions (addendum of arxiv:1611.04135).

Wu, X., Xiao, L., Yixuan, S., Zhang, J., Ma, T., & He, L. (2021). A survey of human-in-the-loop for machine learning.

Xu, S., Li, Y., Liu, S., Fu, Z., Chen, X., & Zhang, Y. (2020). Learning post-hoc causal explanations for recommendation. https://doi.org/10.48550/ARXIV.2006.16977

Yalcin, O., Fan, X., & Liu, S. (2021). Evaluating the correctness of explainable ai algorithms for classification. https://doi.org/10.48550/ARXIV.2105.09740

Yang, M., & Kim, B. (2019). Benchmarking attribution methods with relative feature importance. https://doi.org/10.48550/ARXIV.1907.09701

Yang, Y.-Y., Chou, C.-N., & Chaudhuri, K. (2022). Understanding rare spurious correlations in neural networks. https://doi.org/10.48550/ARXIV.2202.05189

Yang, Z. (2019). Fidelity: A property of deep neural networks to measure the trustworthiness of prediction results. https://doi.org/10.1145/3321705.3331005

Zagalsky, A., Te'Eni, D., Yahav, I., Schwartz, D. G., Silverman, G., Cohen, D., Mann, Y., & Lewinsky, D. (2021). The design of reciprocal learning between human and artificial intelligence. *Proceed-*

ings of the ACM on Human-Computer Interaction, *5*(CSCW2), 1–36. https://doi.org/10.1145/3479587

Zhang, K., Wang, Z., Zhang, J., & Schölkopf, B. (2015). On estimation of functional causal models: General results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *7*(2), 1–22.

Zhang, L., Sun, X., Li, Y., Zhang, Z., & Feng, Y. (2019). A noise-sensitivity-analysis-based test prioritization technique for deep neural networks.