Proceedings of the Tenth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2022)

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 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 in terms 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.

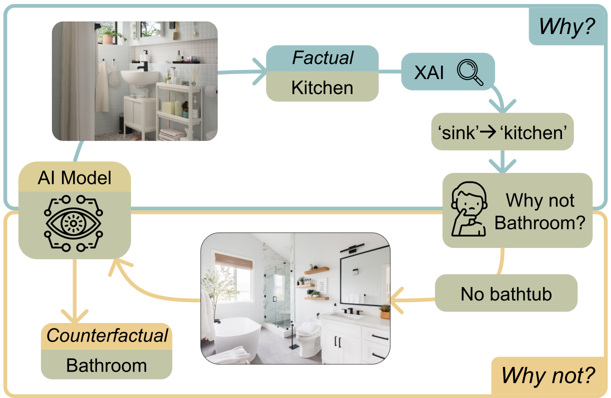


Figure 1: Intuition behind CHIME: to better describe model behaviour, explanations should cover both the factual outcome (i.e., why something occurred) and the hypothetical counterfactual outcome (i.e., why not something different).

Several explanation methods have already been proposed but they only focus on one of the two aspects argued by 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 2014; 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 outcomes 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 their properties 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 due to individual worker factors. In this sense, we analyse the 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 inspecting the causal behaviour of two computer vision models - Inception V3 (Szegedy et al. 2016) 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 dataset are openly available[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 the 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).

Furthermore, they also highlighted the relevant aspect and structure of an explanation. For the most part, past works

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

in philosophy of science and social sciences are critical of 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.

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 one). 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 2014). Saliency can be computed by computing the gradient of the activation functions (Selvaraju et al. 2019) (Simonyan, Vedaldi, and Zisserman 2014), 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. (2018) 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, and the terminology used throughout 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, and 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 2019), or by leveraging asymmetries (Daniušis et al. 2010).

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 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.

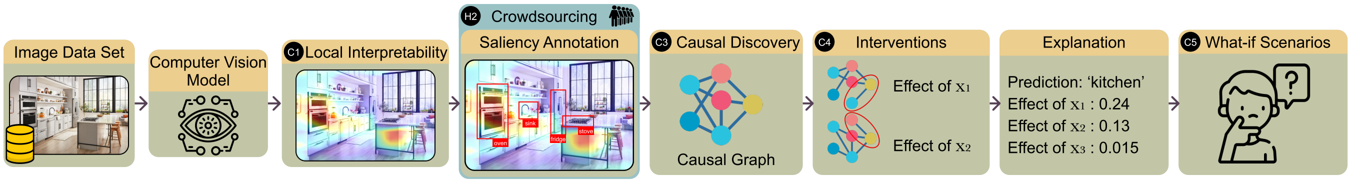


Figure 2: Overview of the CHIME workflow.

H2: Human Annotation Salient patches, in a vacuum, can refer to different concepts: an object, its shape, or its colour. There is no straightforward way to distinguish these individual concepts. Automatic object detection methods 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 shape). We account for the mediating concepts' effects 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. The suggestions were static. Colours were sourced from W3C's Basic Colours list[2]. Shapes and parts of objects are based on commonsense knowledge: squares, circles, handle (of the kitchen utensils), etc.

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 structures with respect to the collected observational data. In our experiments, we showcase and discuss both techniques in the context of explaining model behaviour.

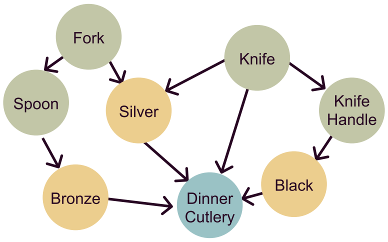


Figure 3: Causal Graph example, including primary concepts (green), mediating concepts (yellow), and the model outcome (blue) for the "Dinner Cutlery" class.

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 and shape in the image, both of which can affect the model outcome as well. Figure 3 shows a simplified example from the causal graphs with our method.

Pairwise Causal Discovery As previously discussed, 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 (2019), given the discreteness of the collected annotations.

C4: Determining Causal Effects Once the causal graph is constructed, we have an overview of the hypothetical model behaviour. However, 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 sin-

[2] https://www.w3.org/wiki/CSS/Properties/color/keywords

gle concept, $do(X)$ is the action of setting the concept X to a particular value, and Z is the set of confounders conditioning the estimates that one needs to account for to not obtain distorted associations. We intervene 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 has on the outcome. Furthermore, this enables us to rank concepts based on the magnitude of their causal effects. Practically, we conduct linear regression on crowd-sourced concepts and 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 → bedroom", and another for "colour → bedroom", where the object is a confounder, i.e., "object → colour" and "object → 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 → 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 → bedroom", and the second one with both object and colour, i.e. "object, colour → 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 2021). 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. In our experiments, we estimate Fidelity by looking at the concepts tagged by crowd workers, which in turn rely on saliency maps. Concretely, we do this 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. 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) 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 trust, 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 causally strongest concepts for each class and fetching 10 images for each concept (50 images in total).

Bias Injection

Sampling Bias (Sackett 1979), a form of collider bias[3], can be injected into models by building ad-hoc datasets such that certain classes are associated with specific, controlled features (e.g., background colour). For example, if considered

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

in a vacuum, the object "knife" has no relation to the colour "black" but, if sampling bias were to be introduced, we create a distorted association between "knife" and "black". This behaviour has also been showcased by 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 photo of a bee as an example: besides the bee itself, the photo may contain other concepts like "flower" or "leaf" which are not attributed by the supervised label *bee*. Furthermore, 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 a model to correctly predict *bee* but for the wrong reasons, i.e., the distracting concepts in the image.

Causality Verification

Inspired by the idea presented by Xu et al. (2020), we check if top causal concepts are indeed causally relevant for a particular class compared to non-causal ones (i.e., with smaller effects). 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 on the primary concept.

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 based on the concepts provided by crowd workers, e.g., given two images, if both of them contain "Spoon" then we consider those two images as similar. 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 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 are accounted for. We perform experiments for both models, on each of the combinations of Causal Discovery strategies and relation types.

Models & Datasets

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

Datasets We consider two datasets: the Edinburgh Kitchen Utensil Database[4] (also referred to as "Utensils"), and ImageNet-A (Hendrycks et al. 2021). The Utensils

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

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

Table 1: Sampling Biases for the Kitchen Utensils dataset.

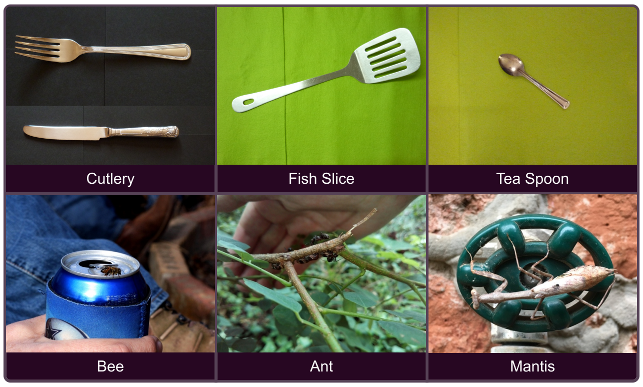


Figure 4: Example images from the Kitchen Utensils dataset (first row), and from ImageNet-A (second row).

dataset contains images of single objects, on solid backgrounds (e.g., completely green), while ImageNet-A contains naturally occurring adversarial images. With the Utensils dataset, we focus on the "Dinner Cutlery"[5], "Fish Slice", and "Tea Spoon" classes while injecting Sampling Bias and controlled noise (summarised in Table 1) based on past work by Zhang et al. (2019) that shows how neural networks are sensitive to noise. To implement this, we insert a few images that are strikingly different (e.g. blue background in a class that is mostly associated with green background). 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 computer vision models. Thus, 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 samples from the 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[6] 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 first explore the efficacy of Template-based and Pairwise Causal Discovery strategies. We consider the 5 concepts having the strongest effects, and compute Kendall's Tau coefficient between those obtained with Template-based and those with 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 the bias (Colour or Shape) introduced. On the other hand, Inception V3 exhibits less stable behaviour with different biases being responsible for the extracted causal concepts to be fairly different. For "Dinner Cutlery" for example, biasing on colour led to relatively similar concepts but in the opposite order, hence the negative value for Kendall's Tau coefficient. In other instances, we see low or no correlation between the extracted concepts. Conversely, the results show less similarity on ImageNet-A regardless of the model. Indeed, this second dataset contains more complex images which led us to collect sparser annotations. However, it is interesting to note how for Inception V3 both causal discovery strategies show signs of positive correlation by returning similar results for the classes "Ant*" and "Mantis" while finding the exact concepts for "Mantis*", albeit with different strengths. Overall, the extracted concepts reflect the differences in the network architectures, with Inception V3 being the more intricate of the two. These differences may have impacted the saliency map generation and hence conditioned what the participants saw during the experiments.

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 are more ambiguous to define and annotate, and thus less frequently found in data.

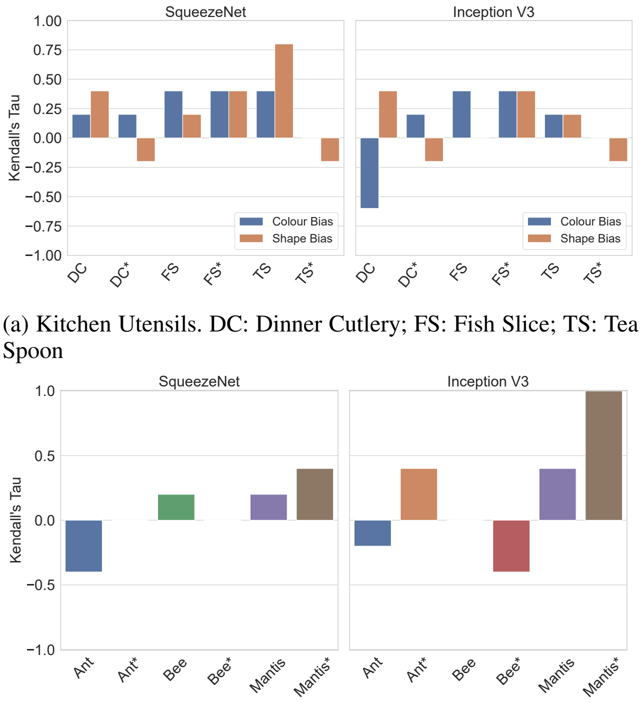
Causality Verification In Table 3 we report the percentage of images that satisfy Inequality 2 aggregated over different combinations of concepts, Causal Discovery strategies (Template-based or Pairwise), and models (Inception V3 or SqueezeNet). Overall, we find concepts extracted for SqueezeNet to score better in terms of Causal Verification with the exception of when only objects and colours are considered ("O-C"). Similarly to Kendall's Tau, such discrep-

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

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

Bias	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_lightbronze** (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)

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.



(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.

ancies may be attributed to the architectural differences between the two models: given our setup and the finite amount of images, Inception V3 may have detected a plethora of non-annotated causal concepts that can influence the Causality Verification estimation. Another factor that might have impacted this evaluation, is the choice of images used, which might also explain the relatively low values we obtained for both models. Finally, 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.

Selection	Template-based		Pairwise	
	Inc. V3	*SN*	*Inc. V3*	*SN*
O - C	52.5±6.3	60.3±10	48.5±19.8	40±13
PO - C	33.3±13.7	46.3±16.5	53.5±13.2	67±12.7
O - C - S	42.5±14.4	60.5±6.3	17.5±11.8	42±6
PO - C - S	43.5±14.9	61±18.7	30.5±13.3	50±6
Average	43±12,3	57±12.9	37.5±14.5	49.8±9.4

Table 3: Causality Verification - average (and standard error) percentages of images satisfying Inequality 2 for different Causal Discovery strategies, Inception V3 (Inc. V3) and SqueezeNet (SN) models, and selections of O) Objects, C) Colours, S) Shapes, and PO) Part-Of Objects.

Mediation Analysis We present in Figure 6 the mediation results from Inception V3 when trained on the colour-biased Utensils dataset. We focus on the "Dinner Cutlery" class. The size of the circles represents the total causal effect of Primary Concepts whereas colours indicate the strengths of Mediating Concepts. We can see that, 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).

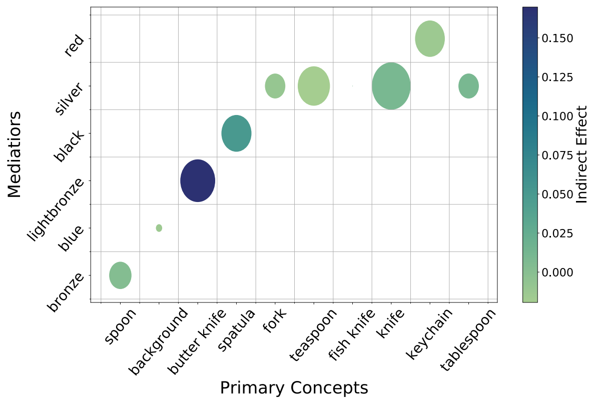


Figure 6: Colour map representing the Mediators' effects on Primary Concepts. The size of the circles represents the sum of Direct and Indirect effects.

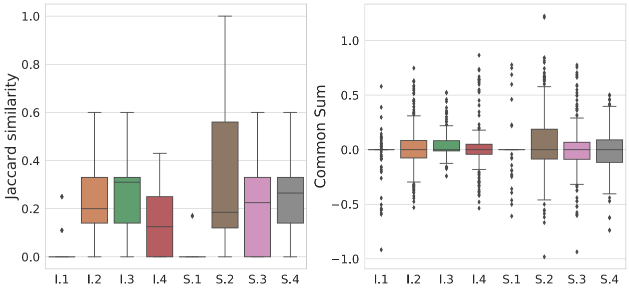


Figure 7: Coherence - Left: Results from Equation 6, Right: Results from Equation 7. The x-axis labels represent different configurations; (I) Inception V3, and (S) SqueezeNet; Biases: (1) Negative Set Bias, (2) No Bias (Utensils), (3) Colour (Utensils), (4) Shape (Utensils).

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 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 applying Causal Inference to XAI. As both are still active research areas, their intersection deserves further analysis. On the other hand, due to the 1:1 relationship between images and annotators present in our design, the possibly inconsistent behaviour of the participants could have made some of the results serendipitous. Furthermore, 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.

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 @2 over 3 classes for different selections of O) Objects, C) Colours, S) Shapes, and PO) Part-Of Objects.

Finally, future work will focus on the scalability of our approach. While annotations can be obtained fairly easily through crowdsourcing, their validation and reconciliation can be improved for larger datasets with more classes.

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 are not easily reduced to linear formulations, 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 models' behaviour.

Acknowledgments

We would like to thank Gaole He for the helpful discussions on the early drafts of this manuscript. This work was partially supported by the TU Delft Design@Scale AI Lab and TU Delft Digital Ethics Centre. In our experiments, we made use of the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-3132. Finally, we would like to thank all participants from Prolific.

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.; 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.

Daniušis, P.; Janzing, D.; Mooij, J.; Zscheischler, J.; Steudel, B.; Zhang, K.; and Schölkopf, B. 2010. Inferring Deterministic Causal Relations. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, 143–150. Arlington, Virginia, USA: AUAI Press. ISBN 9780974903965.

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. 2019. *Conditional Distribution Variability Measures for Causality Detection*, 339–347. Cham: Springer International Publishing. ISBN 978-3-030-21810-2.

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. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2376–2384. PMLR.

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.; Moskewicz, M. W.; Ashraf, K.; Han, S.; Dally, W. J.; and Keutzer, K. 2016. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and ¡1MB model size. *Clinical Orthopaedics and Related Research*, abs/1602.07360.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C. J.; Wexler, J.; Viégas, F. B.; and Sayres, R. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In Dy, J. G.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm,* *Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, 2673–2682. PMLR.

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. *CoRR*, abs/2202.01602.

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. 2021. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.*, 11(3-4): 24:1–24:45.

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. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*.

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. B.; and Wattenberg, M. 2017. SmoothGrad: removing noise by adding noise. *CoRR*, abs/1706.03825.

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. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2818–2826. IEEE Computer Society.

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

van der Waa, J.; Robeer, M.; van Diggelen, J.; Brinkhuis, M. J. S.; and Neerincx, M. A. 2018. Contrastive Explanations with Local Foil Trees. *CoRR*, abs/1806.07470.

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.; and Zhang, Y. 2020. Learning Post-Hoc Causal Explanations for Recommendation. *CoRR*, abs/2006.16977.

Yalcin, O.; Fan, X.; and Liu, S. 2021. Evaluating the Correctness of Explainable AI Algorithms for Classification. *CoRR*, abs/2105.09740.

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.; and Zhang, Z. 2019. A Noise-Sensitivity-Analysis-Based Test Prioritization Technique for Deep Neural Networks. *CoRR*, abs/1901.00054.