Causal Concept Embedding Models: Beyond Causal Opacity in Deep Learning

Gabriele Dominici*
Università della Svizzera italiana
gabriele.dominici@usi.ch

Pietro Barbiero*
Università della Svizzera italiana
pietro.barbiero@usi.ch

Mateo Espinosa Zarlenga
University of Cambridge
me466@cam.ac.uk

Alberto Termine
IDSIA
alberto.termine@idsia.ch

Martin Gjoreski
Università della Svizzera italiana
martin.gjoreski@usi.ch

Giuseppe Marra
KU Leuven
giuseppe.marra@kuleuven.be

Marc Langheinrich
Università della Svizzera italiana
marc.langheinrich@usi.ch

Abstract

Causal opacity denotes the difficulty in understanding the "hidden" *causal structure* underlying a deep neural network's (DNN) reasoning. This leads to the inability to rely on and verify state-of-the-art DNN-based systems especially in high-stakes scenarios. For this reason, causal opacity represents a key open challenge at the intersection of deep learning, interpretability, and causality. This work addresses this gap by introducing Causal Concept Embedding Models (Causal CEMs), a class of interpretable models whose decision-making process is causally transparent by design. The results of our experiments show that Causal CEMs can: (i) match the generalization performance of causally-opaque models, (ii) support the analysis of interventional and counterfactual scenarios, thereby improving the model's causal interpretability and supporting the effective verification of its reliability and fairness, and (iii) enable human-in-the-loop corrections to mispredicted intermediate reasoning steps, boosting not just downstream accuracy after corrections but also accuracy of the explanation provided for a specific instance.

1 Introduction

Deep Learning (DL) models have a pervasive impact on many areas of contemporary research and society [1]. Despite this success, there is growing concern about the widespread real-world application of DL, particularly in sensitive domains [2]. These concerns are partly due to the lack of *causal explainability* of these models, which undermine their robustness, fairness and generalisability [3]. Causal explainability, in particular, is a multi-faced issue, which includes a variety of different, albeit related, problems [4]. One such problem is that of *causal discovery* and concerns the possibility of using a model to detect and understand the *causal mechanisms* of the data generating process [5, 6, 3, 7] (see Figure 1a). An equally important problem is that of *causal opacity*, which denotes the

---

*Equal contribution
Preprint. Under review.

| (a) Causal explainability | (b) Black Box | (c) CBMs | (d) Causal CEMs |

Figure 1: (a) **Causal explainability** has two distinct dimensions: **causal discovery**, where we aim to identify and understand the causal mechanisms underlying a data-generating process, and **causal opacity**, where we aim to understand the causal structure of a model's inference to verify its robustness, generalisability, and fairness. (b) Standard DL models are *black boxes* in the sense that the causal structure of their mapping from raw input features (e.g., pixels of an image) to the target remains opaque. (c) In Concept Bottleneck Models (CBM), high-level human-interpretable concepts are first extracted through an encoder g and then used to predict the target. Although CBMs are semantically interpretable, the causal structure of the model's inference assumes a straightforward causal structure where concepts are causally independent and are all direct causes of the target. (d) In **Causal Concept Embedding Models (Causal CEMs)**, both the concepts' semantics and the inference's causal structure are transparent and interpretable.

difficulty of users to grasp and understand the "hidden" *causal structure* characterising a model's inference and behaviour (see Figure 1b).

Causal opacity can be better understood in the light of Pearl's framework of causality [8, 6], which measures an agent's causal understanding in terms of their ability to answer *what-if* type of questions, and specifically *interventional* and *counterfactual* questions. In this regard, causal opacity depends on the ability of users to answer interventional and counterfactual questions regarding the structure of a model's inference (e.g., "what happens if I fix the feature *age* to a value greater than 50?"). This capacity is vital for assessing a model's reliability and robustness and plays a central role in establishing its fairness, especially within the popular framework of *counterfactual fairness* [9]. For instance, in a loan recommendation system, ensuring fairness might involve verifying that sensitive demographic attributes, like *gender* or *ethnicity*, have no causal influence on model's decision-making process. This task can be solved by checking that such predictions are robust (i.e., do not vary) under interventional and counterfactual adjustments. This ultimately requires answering interventional and counterfactual questions such as "will the decision change if I vary the applicant's ethnicity?".

To address causal opacity, the field of *eXplainable AI* (XAI) has developed a variety of techniques aimed at explaining the causal mechanisms underlying a model's predictions [10–12]. Among the most common methods are *feature attribution techniques*, where an explanation aims to measure the causal relevance of each feature (or latent factor) for a model's outcome [13–20]. Although these techniques are promising, their focus on "raw" input features limits their applicability, especially on unstructured data (e.g., images) whose input features denote low-level attributes (e.g., pixels) lacking an understandable and contextually relevant meaning. For this reason, many works stress that good causal explanations should be *concept-based* [21], i.e., they should rely on the "high-level" features, or "*concepts*", that a model infers from raw data when making predictions [3, 22]. State-of-the-art concept-based models—such as Concept Bottleneck Models [23] and concept-based counterfactuals [24, 25]—partially address this issue while being prone to a key limitation : they only capture *direct counterfactual dependence* [26], a weak form of causal dependence that assumes all concepts to be causally mutually independent and directly related to the target prediction (see Figure 1c). This oversimplifies an inference's causal structure, which typically encompasses a complex, dense network of causal dependencies among various concepts. As a result, the problem of *causal opacity* still represents a key open challenge at the intersection of DL, causality, and XAI that currently limits the explainability, reliability, and verifiability of modern DL systems.

To bridge this gap, we introduce *Causal Concept Embedding Models* (Causal CEMs, see Figure 1d), an interpretable concept-based architecture delivering inferences whose causal structure is transparent by design. The results of our experiments show that Causal CEMs can: (i) match the generalisation

performance of state-of-the-art causally-opaque models, (ii) support the analysis of interventional and counterfactual scenarios, thereby improving the model's causal interpretability and supporting the effective verification of its reliability and fairness, and (iii) enable human-in-the-loop corrections of mispredicted intermediate reasoning steps, boosting not just downstream accuracy after corrections but also accuracy of the explanation provided for a specific instance.

2 Background

Causal Models A causal model (CM) is a mathematical representation of the mechanisms (rules, laws) that explain how different variables of a given target-system causally influence each other. In computer science and statistics, the standard framework for causal modelling is that of *Structural Causal Models* (SCM) proposed by Pearl [8]. Formally, an SCM M is a triplet $(\mathcal{U}, \mathcal{V}, \mathcal{F})$ where: $\mathcal{U}$ is a set of **exogenous variables** representing latent factors with a causal influence on the modelled target-system; $\mathcal{V}$ is a set of **endogenous variables** representing observable and measurable variables; $\mathcal{F}$ is a set of functions (or *structural equations*) describing the **causal mechanisms**, which determine the values of each endogenous variable $v_i \in \mathcal{V}$ by computing $v_i = f_i(u_i, \mathrm{pa}(v_i))$, where $u_i \in \mathcal{U}$ and $\mathrm{pa}(v_i)$ are, respectively, the set of exogenous and endogenous variables causally affecting $v_i \in \mathcal{V}$. Every SCM can be associated with a directed acyclic graph (DAG) whose nodes represent variables and edges represent direct causal connections. Interventions can be modelled in SCMs through the ***do*-operator** [27, 8] . This operator induces a modification on the model's structure by changing the value of an endogenous variable which is fixed to a value $\kappa \in \mathbb{R}$ by external means, effectively breaking its original causal dependencies within the model. Formally, applying $do(v_i = \kappa)$ to a model $\mathcal{M}$ results in a new model $\mathcal{M}_{v_i=\kappa}$ which is identical to $\mathcal{M}$ except that the equation for v_i is replaced with a constant value κ, removing all arrows into v_i in the model annexed DAG.

Concept-based models Concept-based models [28–34] are interpretable architectures that explain their predictions using high-level units of information (i.e., "concepts"). Most of these approaches can be formulated as a Concept Bottleneck Model (CBM) [23], an architecture where predictions are made by composing (i) a concept encoder $g : X \to C$ that maps samples $\mathbf{x} \in X \subseteq \mathbb{R}^d$ (e.g., pixels) to a set of r concepts $c \in C \subseteq \{0,1\}^r$ (e.g., "red", "round"), and (ii) a task predictor $f : C \to Y$ that maps predicted concepts to a set of l tasks $y \in Y \subseteq \{0,1\}^l$ (e.g., labels "apple" or "tomato"). Each component g_i and f_j denotes the truth degree of the i-th concept and j-th task, respectively. Usually, concept-based models represent a concept c_i using its predicted truth degree $\hat{c}_i \in [0,1]$. This representation, however, might significantly degrade task accuracy when the provided concepts are incomplete [35, 36]. To overcome this issue, *Concept Embedding Models* (CEMs) [36] use high-dimensional embeddings $\hat{\mathbf{c}}_i \in \mathbb{R}^z$ to represent concepts alongside their truth degrees $\hat{c}_i \in [0,1]$.

3 Causal Concept Embedding Models

To address the causal opacity of DL systems, we need a model that can answer both interventional and counterfactual causal queries. In order to enable human verification and control, the causal structure of the model's inference should be based on high-level interpretable concepts (e.g., colours, shapes) as low-level attributes do not provide a controllable semantics (a specific pixel lacks meaningful semantics when it comes to causal queries) [21]. In the absence of causal priors, we also need the model to learn causal dependencies among high-level features from available data without making strong assumptions such as direct counterfactual dependence [26]. Finally, the model should not sacrifice predictive generalisation performance in exchange for causal interpretability. To this end, we introduce Causal Concept Embedding Models (Causal CEMs, Figure 1d), a class of concept-based architectures delivering causally transparent and robust inferences. In this section, we present the Causal CEMs blueprint (Section 3.1) and training process (Section 3.2).

3.1 Blueprint

In the absence of causal priors, Causal CEMs need to learn causal dependencies between high-level features that lead to the model becoming more effective in solving its designated task. As a result, we can formally describe Causal CEMs using a generalised probabilistic graphical model (PGM) that extends a traditional PGM by allowing cycles:

Definition 3.1 (Causal Concept Embedding Model). Given an observed input feature x, a set of $k \in \mathbb{N}$ latent factors $u_i \in \mathcal{U}$ each associated with a high level interpretable variable $v_i \in \mathcal{V}$, a *Causal Concept Embedding Model* is the generalised probabilistic graphical model (PGM) $G = (\mathcal{N}, \mathcal{E})$ with nodes $\mathcal{N} = \{x\} \cup \mathcal{U} \cup \mathcal{V}$ and edges $\mathcal{E} = \{(x, u_i) \mid u_i \in \mathcal{U}\} \cup \{(u_i, v_i) \mid i \in \{1, \cdots, k\}\} \cup \{(v_i, v_j) \mid v_i, v_j \in \mathcal{V}, v_i \neq v_j\}$ which represents the joint conditional distribution $p(v, u \mid x)$:

$$(1)$$

The cyclical nature of Causal CEMs comes from the necessity to model all possible dependencies among variables v_i. Notice that the model is uniquely identified by the set of all conditional probability distributions corresponding to the arrows in the graph. Unfortunately, in generalised PGMs, the model does not easily factorise in terms of such distributions due to the cycles. To deal with cycles while maintaining the independencies induced by the graph structure, we can use an unfolding semantics for cyclical PGMs [37]. This semantics is based on the choice of a "cutset" i.e., a specific set of nodes $\mathcal{Q} \subseteq \mathcal{N}$ in the PGM such that every cycle in the PGM contains at least one node in $\mathcal{Q}$. Intuitively, by unfolding the nodes in the cutset, all cycles are broken leaving us with a standard acyclical PGM. The consistency between the semantics of the original cyclical PGM and the unfolded acyclical PGM is only valid in the limit of infinite unfolding [37]. However, when computing the likelihood of an observed complete set of variables $v_i \in \mathcal{V}$, modelling one single unfolding (i.e., a single transition) suffices for learning the conditional probability distributions among the variables in $\mathcal{V}$, as all the variables become conditionally independent on each other. As a result, we can define a *Dissected Causal CEM* as the one-step unfolding of the Causal CEM in Definition 3.1:

Definition 3.2 (Dissected Causal CEM). Given a Causal CEM, let $\mathcal{V}' = \mathcal{V}$ be the cutset. Then, the dissected Causal CEM $\mathcal{G} = (\mathcal{N} \cup \mathcal{V}', \mathcal{E}_{\mathcal{V}'})$ is an acyclic PGM obtained by extending the generalised PGM by (i) adding a copy of all cutset nodes $\mathcal{V}' = \{v_i \mid v_i \in \mathcal{V}\}$, (ii) adding a new set of edges directed from parents of cutset's nodes to the generated copies $\mathcal{V}'$ i.e., $\mathcal{E}_{\mathcal{V}'} = \{(a, b) \mid (a, b) \in \mathcal{E}, b \in \mathcal{V}'\} \cup \{(a, b) \mid (a, b) \in \mathcal{E}, b \notin \mathcal{V}'\}$, and (iii) defining an initial probability distribution for the new copies $p(v'|u)$ given the latent variables. The resulting PGM, factorised as $p(v, v', u \mid x) = p(v \mid v', u)p(v' \mid u)p(u \mid x)$, is:

$$(2)$$

Remark 3.3. The distribution $p(v_i \mid \mathrm{pa}_{\mathcal{V}'}(v_i), u_i)$ can be associated with a structural causal model $\mathcal{M} = (\{u_i\}, \mathcal{V}' \cup \{v_i\}, \{f_i\})$ with causal mechanism $v_i = f_i(u_i, \mathrm{pa}_{\mathcal{V}'}(v_i))$, where $u_i \in \mathcal{U}$ is an exogenous variable representing latent, uninterpretable information (e.g., noise), $v_i \in \mathcal{V}$ is an endogenous variable representing interpretable, symbolic information, and $f_i : U \times V \to V$ is a function describing the causal mechanism that determines the value of v_i given its parents $\mathrm{pa}_{\mathcal{V}'}(v_i)$. Such dependencies are captured by a graph $\mathcal{G} = (\mathcal{V} \cup \mathcal{V}', \{(a, b) \mid a \in \mathcal{V}', b \in \mathcal{V}\})$, representing all direct causal dependencies between endogenous variables.

Remark 3.4. The structure of dissected Causal CEMs resembles the structure of CBMs as the prediction of each v_i can be traced back to $v'_j, \forall j \neq i$. In this sense, endogenous copies in a Causal CEM play the role of "explaining variables", akin to a CBM's concepts. In contrast, endogenous variables play the role of "explained variables", akin to a CBM's tasks.

We can interpret the factors of a dissected Causal CEM as follows: $p(u \mid x)$ is the **exogenous encoder**, i.e., a deterministic distribution that is parametrised by a neural network $\zeta : X \to U$. In CBMs, this function represents the input encoder. The exogenous encoder ζ generates exogenous variables $u_i \in \mathcal{U}$ mapping raw input features $\mathbf{x}$ (e.g., an image's pixels) to latent embeddings $\hat{\mathbf{u}}_i \in \mathbb{R}^q$, $q \in \mathbb{N}$. In practice, this process mirrors the generation of context vectors in Concept Embedding Models [36]. First, the encoder $\psi : X \to H$ maps raw features to a latent code $\mathbf{h} \in H$. Then, a pair of neural networks $\{\phi_i^+, \phi_i^-\}$ map the latent code into two different embeddings whose

concatenation $[\phi_i^+(\mathbf{h}), \phi_i^-(\mathbf{h})]^T$ corresponds to the exogenous variable $\mathcal{U}_i$ of the i-th concept:

$$(\text{exogenous variables}) \quad \hat{\mathbf{u}}_i = \zeta(\mathbf{x}) = [\phi_i^+(\psi(\mathbf{x})), \phi_i^-(\psi(\mathbf{x}))]^T. \tag{3}$$

In contrast, $p(v' \mid u)$ is the **copies predictor**. This is the product of k independent Bernoulli distributions whose logits are parameterised by a neural network $s : U \to V$. In CBMs, the composition of the exogenous encoder and the concept predictor is often called *concept encoder* $g = \zeta \circ s$. In causal methods, this function represents a (supervised) *causal feature learner* [7]. The copies predictor s generates endogenous copies $v_i' \in \mathcal{V}$ from latent embeddings $\hat{\mathbf{u}}_i$. This is obtained using a neural network classifier $s : U \to V$ as a scoring function as in [36]:

$$(\text{endogenous copies}) \quad \hat{v}_i' = s(\hat{\mathbf{u}}_i) = \sigma(W_s \hat{\mathbf{u}}_i + \mathbf{b}_s) \tag{4}$$

Finally, $p(v_i \mid \mathrm{pa}_{\mathcal{V}'}(v_i), u_i)$ is the **endogenous predictor**. This distribution is the product of k independent Bernoulli distributions whose logits are parameterised by a neural network $f : V^k \times U \to V$. The input to this function $\mathrm{pa}_{\mathcal{V}'}(v_i)$ (representing direct causal dependencies) is weighted by a learnable adjacency matrix $M \subseteq \mathbb{R}^{k \times k}$, where each learnable weight m_{ij} models the strength of the dependency of v_i from its parents v_j'. In CBMs this function is called a *task predictor* [23] and intuitively represents the analogous of the structural equations that model *causal mechanisms* in SCMs [7]. The endogenous predictor generates the endogenous variables $\hat{v}_i \in \mathcal{V}$ by considering exogenous variables $\hat{\mathbf{u}}_j$ and copies $\hat{v}_j'$ with $j \neq i$. First, the function $\omega : V \times U \to \mathbb{R}^z$ generates endogenous embeddings $\hat{\mathbf{v}}_j'$ using exogenous variables $\hat{\mathbf{u}}_j$ and copies $\hat{v}_j'$, following [36]. Then, all endogenous embeddings are weighted by the strength of the dependency m_{ij} and aggregated using a deepset-like neural network $f_i : \mathbb{R}^{z \times k} \to [0, 1]$ which maps endogenous embeddings to endogenous predictions:

$$(\text{endogenous embeddings}) \quad \hat{\mathbf{v}}_j' = \omega(\hat{v}_j', \hat{\mathbf{u}}_j) = \hat{v}_j' \phi_j^+(\psi(\mathbf{x})) + (1 - \hat{v}_j')\phi_j^-(\psi(\mathbf{x})) \tag{5}$$

$$(\text{endogenous variables}) \quad \hat{v}_i = f_i\left(\{m_{ij}\hat{\mathbf{v}}_j'\}_{j \in \{1,\ldots,k\}}\right). \tag{6}$$

In order to learn explicit structural equations, existing logic-based aggregation methods can be used from the concept literature [38, 33]. App. A describes in more detail their adaptation in Causal CEM.

3.2 Training Causal CEMs

Learning causal structures Causal CEMs initialise the weights of causal dependencies m_{ij} based on the conditional entropy between labels v_i and v_j, as it can be used to represent asymmetric concept relationship (for other initialisation strategies, see App. A). These weights are then fine-tuned through an end-to-end learning process within the Causal CEM framework following common causal priors, where causal graphs are assumed to be sparse, directed, and acyclic (forming a Directed Acyclic Graph or DAG). We introduce a parameter $\gamma \in \mathbb{R}$ to eliminate less significant dependencies and a loss function, as described by Yang et al. [39], to enforce the sparsity and acyclicity

Figure 2: Unfolded Causal CEM's endogenous predictor.

of the causal graph, ensuring that the adjacency matrix A effectively represents a DAG:

$$(\text{initialization}) \quad m_{ij} = -\sum_{a \in \{0,1\}} \sum_{b \in \{0,1\}} p(v_i = a, v_j = b) \log p(v_i = a \mid v_j = b) \tag{7}$$

$$(\text{sparsity}) \quad A = M \cdot \mathbb{1}_{M \geq \gamma} \tag{8}$$

$$(\text{acyclicity}) \quad \mathcal{L}_2(A) = \mathrm{Tr}\left(\left(I + \frac{\beta}{k} A \cdot A\right)^k\right) - k \tag{9}$$

where k is the number of endogenous, $\mathbb{1}$ an indicator function, and $\beta > 0$ a scaling hyperparameter. *Remark* 3.5 (Unfolding Causal CEMs with directed message passing). Notice how learning a DAG together with Definition 3.2 allows to unfold a Causal CEM's endogenous predictor applying a directed message passing on the associated structural causal model $\mathcal{M}$, ensuring that the values of endogenous variables are derived solely from the nodes that are their ancestors on the causal graph (see Figure 2). As a first step, we compute the exogenous variables for all nodes. We then predict the

values of endogenous variables in root nodes in the learned DAG from their corresponding exogenous variables. Following this, we can generate the endogenous embeddings for root nodes and aggregate endogenous embeddings to compute the value of endogenous variables of each child node. We repeat this process until all leaf nodes of the graph are reached. We can obtain this by replacing in Eq. 5 the endogenous copies $\hat{v}'_j$ with the parents of the endogenous variable v_i:

$$(\textit{unfolding}) \quad \hat{v}_i = f_i\left(\{a_{ij}\omega(\hat{v}_j, \hat{\mathbf{u}}_j)\}\right), \quad \forall i, j \in \mathcal{V} \tag{10}$$

Note that this causal unrolling guarantees two key properties (see Figure 3a): (1) modifying the value of a cause (parent node) will impact the effect (child node) in our model, (2) conversely, intervening upon an effect does not alter the cause. This is because, in our model, information flows sequentially following the graph, mirroring the fundamental nature of causal effects. Consequently, this layer not only facilitates the computation of task predictions but also enables the exploration of causal relationships through do-interventions and counterfactual analysis.

Optimisation problem We can now state the general learning objective for Causal CEMs. Given (1) a set of entities represented by their feature vectors $\mathbf{x} \in \mathcal{X} \subseteq X$ (i.e. *the input*) and (2) a set of annotations for each exogenous variable $v \in \mathcal{V} \subseteq V$ (i.e. *the labels*), we wish to find functions ζ, s, f, together with the adjacency matrix A, that maximise the log-likelihood of v, v', while observing x (or equivalently $u = \zeta(x)$):

$$\mathcal{L} = \overbrace{\mathbb{E}_{u,v' \sim p(u,v')}\left[-\log p(v' \mid u)\right]}^{\text{endogenous copies' prediction}} + \lambda_1 \overbrace{\mathbb{E}_{v \sim p(v|do(v'=v),u)}\left[-\log p(v \mid do(v'=v), u)\right]}^{\text{endogenous variables' prediction}} + \lambda_2 \overbrace{\mathcal{L}_2(A)}^{\text{graph priors}}$$

where $\lambda_{1,2}$ are hyperparameters balancing optimisation objectives. Notice that we use the *do*-operator replacing $\hat{v}'_j$ with labels v_j to minimise leakage and provide better gradients to the endogenous predictor f. This enables Causal CEMs to be aware of *do*-operations during training, thus making the model effective in responding to *do*-interventions once deployed.

4 Causal reasoning and verification with Causal CEM

As traditional CBMs, Causal CEMs enable *ground-truth intervention* that allow domain experts to fix mispredicted concept labels at test time. This notion of interventions must not be confused with the notion of intervention usually discussed in causality literature and modelled by Pearl's *do*-operator. For the latter, here we use the term *do-intervention*. However, in the proposed architecture, the ground-truth interventions have a potentially higher impact as each intervention has a downstream effect on all descendant nodes in the causal graph. In contrast, a concept intervention in a traditional CBM has only an effect on 1-hop nodes representing the downstream task. Moreover, Causal CEMs enable sound causal inference and verification via *do-interventions* and counterfactual analysis.

Ground-truth interventions Causal CEMs support "ground-truth interventions" (see Figure 3b). Ground-truth interventions are one of the core motivations behind CBMs [23]. Through ground-truth interventions, concept bottleneck models allow experts to improve a CBM's task performance by rectifying mispredicted concepts at test time, thus significantly improving task performance within a human-in-the-loop setting. In Causal CEMs, however, ground-truth interventions have a potential impact on all endogenous variables descendant of an intervened node, which may include not only nodes corresponding to downstream tasks but also nodes corresponding to a CBM's intermediate concepts. This enables a single concept ground-truth intervention to potentially improve the prediction of intermediate concepts as well as downstream tasks.

Causal reasoning and verification: do-interventions, counterfactuals, and blocking Causal CEMs can answer interventional and counterfactual queries related to the model's decision-making process using the do-operator on the unfolded SCM. **Do-interventions** enable manipulation of a Causal CEM's decision-making process by changing the value of a specific endogenous variable and observing how it affects other variables' distributions (see Figure 3c). In Causal CEMs, the effect of the *do*-intervention is analysed through the interventional distribution, denoted as $p(v_i|do(v_j = \kappa), \text{pa}(v_i))$, which describes the distribution of the outcome variable v_i after the intervention $do(v_j = \kappa)$ has been performed. In particular, in Causal CEMs, the do-operation fixes the value of the intervened variable v_i to a fixed constant $\kappa \in \{0, 1\}$ and removes all causal dependencies from parent

(a) Causal graph.　　(b) Ground-truth intervention.　　(c) *do* intervention.　　(d) Blocking.

Figure 3: (a) A 5-variable causal graph. (b) A ground-truth intervention fixes the error of the prediction $\hat{v}_3$ to the ground-truth label v_3. (c) A do-intervention sets the value of the second variable to a constant i.e., $v_2 = 0$. The intervention impacts v_2's *effects* i.e., $v_{3,5}$, but does not alter v_2's *causes* i.e., v_1. This operation can override ground-truth interventions. (c) A do-intervention on v_3 *blocks* the causal effects of v_2 on v_5. As a result, intervening on v_2 cannot alter v_5 anymore.

variables by zeroing all values of the i-th column of the adjacency matrix:

$$\text{do}(v_j = \kappa) := \begin{cases} v_j := \kappa, & \kappa \in \{0, 1\} \\ a_{[:,j]} = 0, & (\text{implies that: } \text{pa}(v_j) = \emptyset) \end{cases} \tag{11}$$

Causal CEMs also enable to answer **counterfactual** queries such as "What would the value of the i-th variable have been, had the j-th variable been κ, given that we observed v_i and v_j?". Answering these queries involves three steps [27]: 1) *Abduction*: Infer a realisation of exogenous variables that is consistent with the observed v_i and v_j in the actual causal model. 2) *Action*: Modify the architecture of the causal CEM $\mathcal{M}$ into $\mathcal{M}_{v_j=\kappa}$ by replacing the structural equation for v_j with κ, to simulate the intervention. 3) *Prediction*: Compute the value of v_i in the modified model $\mathcal{M}_\kappa$, representing the counterfactual outcome:

$$(abduction) \quad \hat{\mathbf{u}}_i = \zeta(\mathbf{x}) \tag{12}$$

$$(action) \quad \text{do}(v_j = \kappa) := \begin{cases} v_j := \kappa \\ a_{[:,j]} = 0 \end{cases} \tag{13}$$

$$(prediction) \quad \hat{v}_i = f_i\left(\{a_{in}\omega(\hat{v}_n, \hat{\mathbf{u}}_n)\}\right), \quad \forall i, n \in \mathcal{V} \tag{14}$$

This formalism allows us to not only estimate the effects of hypothetical interventions but also to explore the implications of alternative scenarios on individual outcomes, providing a powerful tool for analysing the model's decision-making based on interpretable causal structures.

Model verification and blocking Causal analysis enables the verification of properties of Causal CEMs before deployment. For instance, using only the learnt causal graph, it is possible to prove that an endogenous variable v_i is independent of the variable v_j by verifying that v_j is not among the ancestors of v_i. Another form of formal verification, which we call "blocking", employs the *do*-intervention (see Figure 3d). Blocking allows one to formally verify the independence of a pair of variables given a sequence of *do*-operations. Given a pair of variables v_j and v_i such that v_j is an ancestor of v_i, we perform a blocking verification as follows: 1) *Block*: perform a do-intervention on all child nodes of v_j, 2) *Verify*: perform a do-operation on v_j itself and observe the impact on v_i. We can easily verify that the first step makes v_j and v_i completely independent by observing that the do-operation on v_j no longer alters the distribution of v_j.

5　Experiments

Our experiments aim to answer the following questions:

- **Concept-based performance and interpretability:** Can Causal CEMs match the generalisation performance of equivalent black-box models and existing CBMs? Can Causal CEMs enable more effective ground-truth interventions w.r.t. existing CBMs?

- **Causal Interpretability:** Are Causal CEMs causally interpretable? Can Causal CEMs effectively block the causal effect of two causally related endogenous variables?

To answer these questions, we use three datasets: (i) Checkmark, a synthetic dataset composed of four endogenous variables; (ii) dSprites, where endogenous variables correspond to object types together with their position, colour, and shape; and (iii) CelebA, a facial recognition dataset where endogenous

variables represent facial attributes. Using these datasets, we compare the proposed approach with a **black box** baseline and state-of-the-art concept-based architectures: Concept Bottleneck Models (**CBM**) [23], and Concept Embedding Models (**CEM**) [36]. We also compare a version of the proposed method (**Causal CEM**) where the causal graph is learnt end-to-end w.r.t. with a version (**Causal CEM+CD**) where the causal graph is extracted from ground-truth labels using a causal discovery algorithm [40]. We provide further details on our experimental setup and baselines in App. B.

A comprehensive set of experiments is detailed in App. C, where the experiments presented in this section for a subset of the datasets are extended to include all datasets.

5.1 Key findings

Causal CEMs match the performance of causally opaque models. (Table 1) Causal CEMs demonstrate robust generalisation across all datasets, yielding a predictive performance close to that of black-box architectures with an equivalent capacity. Causal CEMs using a pre-trained causal graph (Causal CEM+CD) tend to have slightly better label accuracy with respect to Causal CEMs

Table 1: Label accuracy ($\uparrow$) is computed on all endogenous variables (concepts and task).

	CHECKMARK	DSPRITES	CELEBA
Black box	$90.15_{\pm 1.30}$	$99.53_{\pm 0.05}$	$79.55_{\pm 0.14}$
CBM	$90.34_{\pm 0.55}$	$99.55_{\pm 0.07}$	$79.00_{\pm 0.18}$
CEM	$89.09_{\pm 1.98}$	$99.48_{\pm 0.07}$	$79.17_{\pm 0.26}$
Causal CEM+CD	$89.43_{\pm 0.93}$	$99.40_{\pm 0.15}$	$78.42_{\pm 0.42}$
Causal CEM	$88.24_{\pm 1.30}$	$99.44_{\pm 0.11}$	$78.23_{\pm 0.45}$

where the causal graph is learned end-to-end (Causal CEM). Causal CEMs's low variance suggests a consistent robustness on weight initializations over multiple training runs.

Ground-truth interventions on Causal CEMs improve both concept and task accuracy as opposed to CBMs (Figure 4) In Causal CEM, the causal graph induces a natural strategy for ground-truth interventions. Indeed, the causal graph narrows down the set of variables to intervene upon: for any given node, we can just fix mispredicted labels of the node's ancestors as intervening on other nodes will not have any impact. This property significantly decreases the required number of interventions to achieve a desired outcome (e.g., to increase a downstream task accuracy), as shown in App. C. Another advantage of Causal CEM consists in the hierarchical nature of inference which allows ground-truth interventions to impact all endogenous variables descendant of an intervened node. In particular, ground-truth interventions may affect not only nodes corresponding to downstream tasks (as in CBM and CEM), but also nodes corresponding to a CBM's intermediate concepts. We experimentally verify this property and its

Figure 4: Impact of ground-truth interventions on non-intervened nodes ($\uparrow$).

impact by calculating—for nodes that were not intervened upon (including both concepts and tasks)—the change in accuracy before and after ground-truth interventions were applied on their ancestors. Our results (Figure 4) show that Causal CEM improves nodes accuracy by ~ 15 percentage points after only 7 ground-truth interventions on CelebA, while CBM and CEM node accuracy remains almost unchanged. Causal CEM's advantage increases with the number of concepts and connections, as a single intervention can impact a higher number of nodes in the causal graph. CBM and CEM, instead, achieve a similar performance only after intervening on all concepts as their architecture assumes all concepts to be mutually independent.

Causal CEMs' endogenous predictors are causally interpretable (Figure 5) In Causal CEM the decision-making process is causally interpretable and can be analysed by visualising the learnt causal graph and structural equations as in a structural causal model, as shown in Figure 5 for CelebA. The image shows how Causal CEM exploited known biases in CelebA to infer facial attributes. For instance, Causal CEM predicts the attribute "wearing lipstick" directly from the attribute "attractive" and indirectly from attributes such as "smiling" or "high cheek", all attributes that are known to be strongly correlated with each other in CelebA [41, 42].

We can also quantify the strength of the causal dependency between two nodes by computing the probability of necessity and sufficiency (PNS) [43]. In the figure, we represent the PNS w.r.t. the leaf node by colouring each node with a different shade of orange. This shows how, for Causal CEM, the attribute "heavy makeup" has the strongest impact on the leaf node. This high degree of causal transparency allows users to interpret Causal CEM's inference and can be eventually exploited to identify potential biases, thereby supporting the assessment of the model's counterfactual fairness. As a result, users can intervene directly on the causal structure of the decision-making process and remove biases using do-interventions, as shown in the next paragraph.

Figure 5: Portion of the learnt causal graph and structural equations in CelebA. A node's colour in the causal graph is proportional to the probability of necessity and sufficiency w.r.t. the node v_5.

Causal CEMs can make two causally-related variables causally independent by blocking all paths between these variables (Table 2) The causal transparency of the proposed approach allows users to modify the model's decision-making process (e.g., to de-bias the model's inference) by using do-interventions. In particular, we can make two causally-related variables i and j causally independent by blocking all paths between the cause i and the effect j. We experimentally verify this property and its impact by computing the Residual Concept Causal Effect i.e., the ratio between the Concept Causal Effect (CaCE) [44] obtained after and

Table 2: Residual Concept Causal Effect ($\downarrow$) between causally-related variables having blocked all paths between the two variables with do-interventions on the causal graph. The optimal value is zero corresponding to perfect causal independence. Values above 100% mean that the causal effect increased instead of decreasing.

	CHECKMARK	DSPRITES	CELEBA
Black box	N/A	N/A	N/A
CBM	$97.99_{\pm 5.64}$	$100.00_{\pm 0.70}$	$97.84_{\pm 2.13}$
CEM	$102.58_{\pm 12.95}$	$100.00_{\pm 4.62}$	$106.00_{\pm 0.50}$
Causal CEM+CD	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$
Causal CEM	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$

before blocking. The optimal value of this metric is zero, corresponding to perfect causal independence between i and j (i.e., the optimal value for a de-biasing operation). The results show that, in Causal CEMs, blocking a variable in the causal graph always yields a perfect Residual Concept Causal Effect of zero across all datasets. In contrast, applying the same procedure in CBMs leads only to a negligible reduction in the average causal effect to 3 percentage points. CEMs not only fail to reduce the causal effect to zero but, in some cases, even increase the causal influence. These results underscore how Causal CEMs transparency enable users to manipulate the model's decision-making process to achieve desired outcomes, as opposed to existing CBMs.

6 Discussion

Related works Causal CEMs present substantial advantages compared to the state of the art. Compared to most causal feature-attribution methods (e.g., [13]), Causal CEMs focus on high-level human interpretable concepts. Causal CEMs differ from existing CBMs in their approach to intervention and causal relationships. The causal structure of the inferences that Causal CEMs deliver does not subsume weak forms of causal dependencies as in existing CBMs, where concepts are all direct causes of the target and causally independent of each other. This way, Causal CEMs enable human-in-the-loop corrections to mispredicted intermediate reasoning steps, boosting not just downstream accuracy after corrections but also the accuracy of the explanation provided for a specific instance. Closest to Causal CEMs in concept usage are *post-hoc* causal concept-based explainability techniques, like DiConStruct [22] and conceptual counterfactual explanations [25]. These methods build surrogate causal models to emulate a target black box model's predictive behavior. However, as Rudin [45] notes, convergence in predictive behavior does not ensure structural similarity in decision-making processes, making surrogate models as explanatory proxies questionable. Causally interpretable *by-design* architectures, such as Causal CEMs, do not suffer from this issue.

Limitations and future works Our method's limitations are mainly derived from limitations inherent to CBMs and causal reasoning. The quality of learnt causal graphs mainly depends on the quality of the dataset and its annotations. Missing and noisy labels might lead to suboptimal graphs.

Similarly to generalised PGMs, Causal CEMs can be easily unfolded when the final graph is acyclic. Variables in cycles can still be inferred, but require special unfolding techniques which might be explored in future works. Finally, a Causal CEM's learnt graph does not necessarily represent the causal mechanisms of the data-generating process, but rather that of the model's inference. This makes our models suitable to be verified and controlled but not necessarily to understand their training data's distribution.

Conclusion and impact   Causal opacity represents a key open challenge at the intersection of deep learning, interpretability, and causality. Causal CEMs address this challenge by employing an architecture which makes the decision-making process causally transparent by design. This makes Causal CEMs reliable and verifiable compared to both usual DL architectures and standard (non-causal) CBMs. The results of our experiments show that Causal CEMs support the analysis of interventional and counterfactual scenarios—thereby improving the model's causal interpretability and supporting the effective verification of its reliability and fairness—and enable human-in-the-loop corrections to mispredicted intermediate reasoning steps, boosting not just downstream accuracy after corrections, but also accuracy of the explanation provided for a specific instance. As a result, advancing this research line could significantly improve the reliability and verifiability of concept-based deep learning models, thus supporting their deployment in real-world applications.

References

[1] Pierre Baldi. *Deep learning in science*. Cambridge University Press, 2021.

[2] Virginia Dignum. *Responsible artificial intelligence: how to develop and use AI in a responsible way*, volume 1. Springer, 2019.

[3] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[4] Alberto Termine and Giuseppe Primiero. Causality problems in machine learning systems. In Federica Russo and Phyllis Illari, editors, *Routledge Handbook of Causality and Causal Methods*. Routledge, 2024 *forthcoming*.

[5] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[6] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

[7] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

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

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

[10] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[11] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[12] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[13] Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective. In *International Conference on Machine Learning*, pages 981–990. PMLR, 2019.

[14] Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33:1229–1239, 2020.

[15] Jiaxuan Wang, Jenna Wiens, and Scott Lundberg. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, pages 721–729. PMLR, 2021.

[16] Tom Heskes, Evi Sijben, Ioan Gabriel Bucur, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789, 2020.

[17] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pages 2907–2916. PMLR, 2020.

[18] Aria Khademi and Vasant Honavar. A causal lens for peeking into black box predictive models: Predictive model interpretation via causal attribution. *arXiv preprint arXiv:2008.00357*, 2020.

[19] Qingyuan Zhao and Trevor Hastie. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281, 2021.

[20] Yonghan Jung, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Blöbaum, and Elias Bareinboim. On measuring causal contributions via do-interventions. In *International Conference on Machine Learning*, pages 10476–10501. PMLR, 2022.

[21] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688, 2019.

[22] Ricardo Miguel de Oliveira Moreira, Jacopo Bono, Mário Cardoso, Pedro Saleiro, Mário AT Figueiredo, and Pedro Bizarro. Diconstruct: Causal concept-based explanations through black-box distillation. In *Causal Learning and Reasoning*, pages 740–768. PMLR, 2024.

[23] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

[24] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[25] Abubakar Abid, Mert Yuksekgonul, and James Zou. Meaningfully explaining model mistakes using conceptual counterfactuals. *arXiv preprint arXiv:2106.12723*, 2021.

[26] Sander Beckers. Causal explanations and xai. In *Conference on causal learning and reasoning*, pages 90–109. PMLR, 2022.

[27] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[28] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[29] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

[30] Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33:20554–20565, 2020.

[31] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

[32] Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. *arXiv preprint arXiv:2306.01574*, 2023.

[33] Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lio, Frederic Precioso, Mateja Jamnik, and Giuseppe Marra. Interpretable neural-symbolic concept reasoning. *arXiv preprint arXiv:2304.14068*, 2023.

[34] Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

[35] Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

[36] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models. *Advances in Neural Information Processing Systems*, 35, 2022.

[37] Christel Baier, Clemens Dubslaff, Holger Hermanns, and Nikolai Käfer. On the foundations of cycles in bayesian networks. In *Principles of Systems Design: Essays Dedicated to Thomas A. Henzinger on the Occasion of His 60th Birthday*, pages 343–363. Springer, 2022.

[38] Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Pietro Lió, Marco Gori, and Stefano Melacci. Entropy-based logic explanations of neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6046–6054, 2022.

[39] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*, 2020.

[40] Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In James Cussens and Kun Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 1052–1062. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/lam22a.html.

[41] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9301–9310, 2021.

[42] Angelina Wang and Olga Russakovsky. Overcoming bias in pretrained models by manipulating the finetuning dataset. *arXiv preprint arXiv:2303.06167*, 2023.

[43] Judea Pearl. Probabilities of causation: three counterfactual interpretations and their identification. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 317–372. 2022.

[44] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace), 2020.

[45] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[46] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[47] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[48] Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models, 2023.

[49] Gabriele Ciravegna, Pietro Barbiero, Francesco Giannini, Marco Gori, Pietro Lió, Marco Maggini, and Stefano Melacci. Logic explained networks. *Artificial Intelligence*, 314:103822, 2023.

[50] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[52] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9 (3):90–95, 2007. doi:10.1109/MCSE.2007.55.

A Architecture

Initialization of adjacency matrix Causal CEM provides versatile initialisation options for the adjacency matrix A, tailored to specific scenarios. In certain instances, weights can be derived from domain expertise or provided along with training data and labels. Without such information, weights can be directly inferred from training labels through causal structural learning algorithms [7] as a preliminary step. These approaches guide the model towards a predetermined decision-making pathway. Alternatively, weights can be learnt concurrently during the Causal CEM training phase, as outlined in Section 3.2. These weights may be initialised either randomly or based on the conditional entropy between labels, providing a better starting point. Additionally, a hybrid approach is feasible, where certain elements in A are fixed while others remain trainable. For instance, a causal structural learning algorithm might yield a Partial Ancestral Graph (PAG) with undirected edges, allowing for the definition of directed edges and learning the direction for others to avoid cycle formation.

Causal mechanisms In Causal CEM, the function f_i corresponds to a causal mechanism in a SCM. Such mechanisms are typically formalized via structural equations. For instance, linear models are a common choice for label predictors in Concept Bottleneck Models [23] where endogenous embeddings are aggregated using a permutation invariant aggregator function $\oplus$ (such as the element-wise maximum, or sum):

$$\hat{v}_i = \sigma \left(W_i \bigoplus_{j \in \{1,\ldots,k\}} m_{ij} \hat{\mathbf{v}}_j + \mathbf{b}_i \right) \tag{15}$$

However, other options are also available to increase the expressiveness and interpretability of the decision-making process, such as Deep Concept Reasoning [33] class predictors, which build logic-based formulae to obtain class label predictions using endogenous embeddings:

$$\hat{v}_i \leftarrow \bigvee_{\mathbf{x} \in X_{\text{train}}} \bigwedge_{j \in \{1,\ldots,k\}} l_j(\mathbf{x}) = \bigvee_{\mathbf{x} \in X_{\text{train}}} \bigwedge_{j \in \{1,\ldots,k\}} (\rho_{ij}(m_{ij}\hat{\mathbf{v}}_j) \iff \hat{v}_j') \tag{16}$$

where l_j denotes the literal of relevant $\hat{v}_j'$ representing the variable's sign or "polarity" in the logic rule (i.e., either v_j or $\neg v_j$). For example, given three variables v_1, v_2, v_3, v_4, DCR can predict the endogenous variable v_2 using the rule $v_2 \leftarrow (v_1 \wedge \neg v_3) \vee (\neg v_1 \wedge v_3)$ which highlights the underlying causal mechanism linking v_1, v_3, v_4 to v_2 (notice how DCR can also learn to remove irrelevant variables such as v_4).

Compositional generalization The training procedure of Causal CEMs is highly parallelizable and modular as only direct connections need to be trained together (e.g., $a \to b$ and $b \to c$), while the model takes care of distant connections in an indirect way. For instance, the connection $a \to b \to c$ can be obtained as a composition of two different independent training procedures for $a \to b$ and $b \to c$. As a result, it is trivial for a Causal CEM to make the causal graph grow even at test time (see Figure 6). This can be done by composing two different graphs obtained by independent training procedures, encoders, datasets, or data types. We note that this is not possible in standard CBMs, which need to re-train the task predictor from scratch whenever new concepts or tasks are added to the mix. Moreover, this modularity enables a form of out-of-distribution compositional generalization as it creates new distant connections between variables that were never part of the same training procedure (e.g., a and c in the previous example).

Figure 6: Compositional generalization in Causal CEMs: two different Causal CEMs architectures are trained independently and then composed only at test time, thus creating a larger graph and allowing out-of-distribution causal inference.

B Experimental setup

B.1 Datasets

In our experiments we use three different datasets:

- **Checkmark** — The dataset consists of tabular data with three features, each ranging from -1 to 1 (denoted as a, b, and c). The target variable d can be either 0 or 1. Each feature is annotated with a concept that indicates whether it is positive or negative. The dataset also incorporates causal relationships among the features. For example, feature c is defined as the inverse of feature b. The target d is set to 1 when both features a and b are positive. This data set is used to test our hypothesis in a straightforward and controlled setting.

- **dSprites** [46] — The dataset comprises images featuring one of three objects (square, heart) in various positions and sizes. The defined concepts include: (1) object shape (square or heart), (2) object size (small or large), (3) vertical position (top or bottom), (4) horizontal position (left or right), (5) object colour (red or blue). Based on these, causal relationships and a binary classification task are established: if the object is a heart on the right side, it is large; if a heart is at the top of the image, it is red; the label is positive if the object is both red and large.

- **CelebA** [47] — The CelebA dataset features celebrity images annotated with various attributes, including lipstick presence, gender, facial shape, and hair type. Gender is used as the classification label. This dataset is chosen for the presence of correlations and biases, such as the association between wearing lipstick and being identified as female.

B.2 Baselines

We evaluated our approach against three established baselines:

- **Black Box**: This model employs a single predictor that processes the input to simultaneously predict the task label and all relevant concepts. It lacks interpretability and does not differentiate between the importance of task labels and concepts.

- **Concept Bottleneck Model (CBM) [23]**: This model first uses a concept predictor to infer concepts from the initial input, followed by a task label prediction based on these concepts. It is designed to be interpretable and treats concepts as significant informational to predict the task label.

- **Concept Embedding Model (CEM) [36]**: Comprising n context encoders, one for each concept, this model predicts each concept based on its respective context before predicting the final task label. It treats concepts and task labels in the same way as CBM.

B.3 Experiments

In our experiments, we evaluate our approach by examining four key dimensions: (1) performance accuracy, (ii) influence of ground-truth interventions, (iii) identification of causal structures, and (iv) blocking for the influence of one variable on another.

To evaluate the first dimension, we conducted a comparative analysis of our approach (using both a learned and a predefined graph) against Black Box, CBM and CEM. This was to determine if graph-based inference would decrease model performance. For this assessment, we calculated the model's accuracy in predicting all concepts and the task. Typically, these metrics are calculated independently; however, in our study, we treated tasks and concepts equivalently, considering them collectively as labels.

In the second aspect, we evaluate our approach by comparing it against CBM and CEM in terms of response to ground-truth interventions. Enhancing the impact of interventions in the Concept-Based Model is crucial for improving the role of humans in the loop. In our experiments, we initially perturbed the inputs to reduce label prediction accuracy, following methodologies established in prior research [36]. Subsequently, we implemented interventions on the most inaccurately predicted concepts in CEM and CBM. This intervention strategy is considered highly effective, as noted in [48]. For the Causal CEM, interventions began with concepts that have a higher number of descendant

nodes in the model's graph, aiming to maximise the intervention's effectiveness. To assess this dimension, we measured the change in accuracy for non-intervened labels before and after the interventions on n concepts (Delta Label Accuracy).

In the third aspect, we visualise the DAG utilised during the inference stage by Casual CEM and derive the corresponding logic equations. We generate Sum of Product logic rules from a table that lists all possible combinations of input concept values alongside the most frequent prediction for each combination derived from the training set, similarly to what done by [49]. It is crucial to note that while these logic rules are general for the model decision-making process, exogenous information may alter predictions for particular instances.

In the final dimension of our analysis, we compare Causal CEM, which operates on a specified graph, against CBM and CEM in terms of their efficacy in mitigating a variable's influence on the task. Specifically, we perturb the prediction of a concept and then, following the graph structure utilised by Causal CEM, we intervene with the ground truth label on all descendant nodes (blocking). Consequently, in Causal CEM, all links between the altered concepts and the labels are deleted, a feat unachievable in CBM and CEM. To assess this characteristic, we calculated the Residual Concept Causal Effect, ratio of the Concept Causal Effect [44] post- and pre-application of the blocking techniques. Ideally, this ratio should be zero, indicating that after blocking, the altered node's value no longer influences the outcome of the task.

B.4 Implementation details

Additional details To maximise the efficacy of interventions in Causal CEM, the second term of the loss can be regularised to maximise the average Causal Concept Effect (CaCE) [44] as follows:

$$\mathcal{L}_{\text{CaCE}} = \frac{1}{n} \sum_{i=0}^{n} |p(v_i|do(v_{i,r} = 1)) - p(v_i|do(v_{i,r} = 0))|$$

Here, r represents a randomly chosen index for each sample, which is used to select one of the concepts following Espinosa Zarlenga et al. [36]. This regularisation can be weighted using an hyperparameter, λ_3. Moreover, for all the experiments where the graph is learnt end-to-end, we initialise the learnable adjacency matrix with the conditional entropy between each pair of values, extracted from the training set.

Hyperparameters All baseline and proposed models were trained for varying epochs across different datasets: 500 for Checkmark, 200 for dSprites, and 30 for CelebA. The optimal epoch for each was determined based on label accuracy on the validation set. A uniform learning rate of 0.01 was applied across all models and datasets. For the CBM and CEM models, both concept and task losses were equally weighted at 1. This weighting scheme was also applied to the loss terms for endogenous copies' prediction, endogenous variables' prediction (λ_1), and graph priors (λ_2). The weight assigned to the loss terms in our models to maximize CaCE is 0.05. Additionally, γ was treated as a learnable parameter, initialized at 0.1, and β was set to 1. All experiments were conducted using five different seeds (1, 2, 3, 4, 5).

Code, licenses and hardware For our experiments, we implement all baselines and methods in Python 3.9 and relied upon open-source libraries such as PyTorch 2.0 [50] (BSD license), PytorchLightning v2.1.2 (Apache Licence 2.0), Sklearn 1.2 [51] (BSD license). In addition, we used Matplotlib [52] 3.7 (BSD license) to produce the plots shown in this paper. Two datasets we used are freely available on the web with licenses: dSprites (Apache 2.0) and CelebA, which is released for non-commercial use research purposes only. We also introduce the Checkmark dataset and we described it in this section. We will publicly release the code with all the details used to reproduce all the experiments under an MIT license. The experiments were performed on a device equipped with an M3 Max and 36GB of RAM, without the use of a GPU. Approximately 40 hours of computational time were utilized from the start of the project, whereas reproducing the experiments detailed here requires only 2 hours.

C Additional results

In this section, we include all the experiments shown in Section 5 for the three datasets in more detail and an ablation study on the value of λ_3.

C.1 Ground-Truth interventions

Figure 7 illustrates the performance comparison among Causal CEM (using both the provided and learnt graphs), CBM, and CEM regarding the effects of interventions. Delta Label Accuracy, which quantifies the change in label accuracy before and after interventions on a growing number of concepts, is calculated solely for the concepts not directly intervened upon. In particular, Causal CEM demonstrates superior performance when interventions involve fewer concepts. This superior performance is attributed to the propagation of intervention effects through all descendant nodes in Causal CEM, unlike CBM and CEM, where the impact is confined to the final task without adjustments to other concepts. The most significant performance gain, approximately 15 percentage points, is observed in CelebA after seven interventions. This effect is particularly pronounced in scenarios with multiple concepts, such as in dSprites and CelebA. However, in simpler tasks with fewer concepts, like Checkmark, the advantage offered by our method is lower. Furthermore, when focusing solely on the effects of interventions on the task label, the causal graph utilised by Causal CEM allows us to identify beforehand the specific subset of concepts influencing the task prediction. This pre-identification significantly decreases the required number of interventions to achieve the desired outcome. Figure 8 demonstrates that Causal CEM attains comparable improvements in task performance but after interventions on only three or four concepts, in contrast to the ten and eleven concepts required by CEM and CBM, respectively. The elevated standard error observed in Causal CEM with the learnt graph is attributed to the variability of the graph structure, which significantly influences the outcomes of interventions.

C.2 Causal structures

Figure 9 illustrates the adjacency matrices corresponding to the DAGs used by Causal CEM for inference in three datasets. On the other hand, Tables 3, 4, and 5 present the logic rules derived from the adjacency matrices depicted in the aforementioned figure. Notably, in the Checkmark dataset, both configurations successfully identified the ground truth graph and the correct logic rules. In the case of dSprites, the DAG identified through causal structural learning (GRaSP [40]) accurately discovers the causal graph and associated logic rules. Although the end-to-end model accurately identifies the correct relationships between concepts and tasks, it proposes alternative methods for concept prediction. It is important to note that even though the model did not identify the correct causal graph, the model was still capable of performing causal inference with the existing graph. In the CelebA data set, where there is no ground truth for either the graph or logic rules, the findings by GRaSP and the end-to-end model appear plausible and reveal biases inherent in the dataset, such as the strong correlation between makeup use and gender or potential causal links like smiling and a slightly open mouth. This scenario underscores the benefits of employing Causal CEM, particularly in demonstrating how specific concepts are used to predict other concepts and tasks.

C.3 Ablation study

In Tables 6, 7, and 8, we present the outcomes of varying the hyperparameter λ_3, which weights the loss term designed to enhance the CaCE effect. The results indicate that optimising this loss term contributes to improved CaCE scores, thereby augmenting the efficacy of the interventions. Nonetheless, excessively high values of λ_3 may lead to diminished model performance, as it tends to prioritise boosting the CaCE score at the expense of accurate predictions.

(a) Checkmark   (b) dSprites   (c) CelebA

Figure 7: Impact of ground-truth interventions on concepts across three datasets. This figure illustrates the variations in accuracy for non-intervened labels, comparing performance before and after interventions on specific nodes.

Figure 8: Impact on the task accuracy of ground-truth interventions performed on CelebA concepts. Causal CEM+CD has received in input a causal graph, discovered with a causal structural learning algorithm (GRaSP [40]), while Causal CEM learns it end-to-end.

Table 3: Logic rules extracted for the Checkmark dataset from Causal CEM+CD with a given DAG and from Causal CEM with a learnt DAG. A term which refers to an exogenous variable is omitted for simplicity.

METHOD	CHECKMARK
Causal CEM	$a \leftarrow \epsilon_0$
	$b \leftarrow \epsilon_1$
	$c \leftarrow \sim b$
	$d \leftarrow a \wedge c$
Causal CEM+CD	$a \leftarrow \epsilon_0$
	$b \leftarrow \epsilon_1$
	$c \leftarrow \sim b$
	$d \leftarrow a \wedge c$

18

(a) Causal CEM with learnt graph   (b) Causal CEM with given graph

(c) Causal CEM with learnt graph   (d) Causal CEM with given graph

(e) Causal CEM with learnt graph   (f) Causal CEM with given graph

Figure 9: Adjacency matrices representing the DAG used by Causal CEM during inference on the three datasets. On the left side, the matrices represent the DAG learnt end-to-end by the model, while on the right the DAG discovered with GRaSP [40]. It provides a PAG starting from the training data.

Table 4: Logic rules extracted for the Dsprites dataset from Causal CEM+CD with a given DAG and from Causal CEM with a learnt DAG. A term which refers to an exogenous variable is omitted for simplicity.

METHOD	DSPRITES
Causal CEM	Shape $\leftarrow$ Size
	Size $\leftarrow \epsilon_1$
	PosY $\leftarrow \epsilon_2$
	PosX $\leftarrow \epsilon_3$
	Color $\leftarrow$ Shape
	Label $\leftarrow$ Size $\wedge$ Color
Causal CEM+CD	Shape $\leftarrow \epsilon_0$
	Size $\leftarrow$ Shape $\wedge$ PosX
	PosY $\leftarrow \epsilon_2$
	PosX $\leftarrow \epsilon_3$
	Color $\leftarrow$ Shape $\wedge$ PosY
	Label $\leftarrow$ Size $\wedge$ Color

Table 5: Logic rules extracted for the Celeba dataset from Causal CEM+CD with a given DAG and from Causal CEM with a learnt DAG. A term which refers to an exogenous variable is omitted for simplicity.

METHOD	CELEBA
Causal CEM	Smiling (S) $\leftarrow \epsilon_0$
	Attractive (A) $\leftarrow$ Heavy_Make
	Mouth_Slig (MS) $\leftarrow$ False
	High_Cheek (HC) $\leftarrow \epsilon_3$
	Wearing_Li (WL) $\leftarrow$ Attractive
	Heavy_Make (HM) $\leftarrow$ Smiling $\wedge$ High_Cheek
	Male $\leftarrow\sim$ Wearing_Li$\wedge \sim$ Heavy_Make
	Wavy_Hair (WH) $\leftarrow$ (HC $\wedge$ WL$\wedge \sim$ Male) $\vee$ (HC $\wedge$ WL$\wedge \sim$ OF)
	Big_Lips (BL) $\leftarrow$ Smiling $\wedge$ High_Cheek$\wedge \sim$ Male$\wedge \sim$ Oval_Face
	Oval_Face (OF) $\leftarrow$ False
	Makeup (M) $\leftarrow$ False
	Fem_Model $\leftarrow$ (M$\wedge \sim$ S) $\vee$ (M$\wedge \sim$ Male) $\vee$ (M $\wedge$ HC$\wedge \sim$ WH) $\vee$ (WL $\wedge$ WH $\wedge$ BL$\wedge \sim$ S$\wedge \sim$ HC)
Causal CEM+CD	Smiling $\leftarrow$ Mouth_Slig
	Attractive $\leftarrow$ Wearing_Li
	Mouth_Slig $\leftarrow \epsilon_2$
	High_Cheek $\leftarrow$ Smiling
	Wearing_Li $\leftarrow \epsilon_4$
	Heavy_Make $\leftarrow$ (Attractive $\wedge$ Wearing_Li) $\vee$ (Wearing_Li $\wedge$ Oval_Face)
	Male $\leftarrow\sim$ Wearing_Li$\wedge \sim$ Heavy_Make
	Wavy_Hair $\leftarrow$ Makeup $\vee$ (Attractive $\wedge$ Wearing_Li$\wedge \sim$ Male)
	Big_Lips $\leftarrow$ Makeup
	Oval_Face $\leftarrow$ Smiling $\wedge$ Attractive $\wedge$ Wearing_Li$\wedge \sim$ Big_Lips
	Makeup $\leftarrow$ False
	Fem_Model $\leftarrow$ Makeup

Table 6: Ablation study regarding λ_3 on the Checkmark dataset.

	λ_3	Label Accuracy	Average CaCE	min CaCE	max CaCE	CaCE	CaCE block
Black Box		90.15 ± 0.14					
CBM		90.34 ± 0.55	4.09 ± 0.80	0.00 ± 0.00	9.45 ± 1.86	28.36 ± 5.58	27.79 ± 5.64
CEM		89.09 ± 1.98	1.75 ± 1.66	0.00 ± 0.00	4.45 ± 3.78	11.59 ± 12.76	11.89 ± 12.95
Causal CEM	0.05	88.24 ± 1.30	14.47 ± 2.86	0.00 ± 0.00	44.37 ± 5.58	16.15 ± 11.25	0.00 ± 0.00
Causal CEM + CE	0.05	89.43 ± 0.93	13.15 ± 2.31	0.00 ± 0.00	39.53 ± 3.62	20.99 ± 8.19	0.00 ± 0.00
Causal CEM	0.2	85.88 ± 3.31	11.64 ± 3.21	0.00 ± 0.00	37.55 ± 8.77	16.32 ± 15.36	0.00 ± 0.00
Causal CEM + CE	0.2	85.09 ± 2.78	16.39 ± 5.00	0.00 ± 0.00	39.69 ± 10.97	31.15 ± 19.48	0.00 ± 0.00
Causal CEM	0	87.86 ± 1.66	6.98 ± 2.95	0.00 ± 0.00	18.16 ± 8.29	7.39 ± 4.10	0.00 ± 0.00
Causal CEM + CE	0	87.04 ± 2.72	12.16 ± 2.92	0.00 ± 0.00	33.15 ± 9.97	14.52 ± 14.29	0.00 ± 0.00

Table 7: Ablation study regarding λ_3 on the dSprites dataset.

	λ_3	Label Accuracy	Average CaCE	min CaCE	max CaCE	CaCE	CaCE block
BlackBox		99.53 ± 0.05					
CBM		99.55 ± 0.07	2.77 ± 0.30	0.00 ± 0.00	8.51 ± 1.03	0.64 ± 0.70	0.64 ± 0.70
CEM		99.48 ± 0.07	0.46 ± 0.29	0.00 ± 0.00	2.45 ± 1.67	2.43 ± 4.62	2.43 ± 4.62
CausalCEM	0.05	99.44 ± 0.11	17.53 ± 3.26	0.00 ± 0.00	44.47 ± 10.16	34.13 ± 19.46	0.00 ± 0.00
Causal CEM + CE	0.05	99.40 ± 0.15	12.85 ± 0.59	0.00 ± 0.00	27.72 ± 3.66	28.95 ± 13.50	0.00 ± 0.00
Causal CEM	0.2	98.80 ± 1.24	14.13 ± 3.39	0.00 ± 0.00	37.59 ± 14.20	17.52 ± 13.41	0.00 ± 0.00
Causal CEM + CE	0.2	99.30 ± 0.13	12.90 ± 0.51	0.00 ± 0.00	34.01 ± 6.99	16.84 ± 6.79	0.00 ± 0.00
Causal CEM	0	99.58 ± 0.12	6.89 ± 1.55	0.00 ± 0.00	18.51 ± 5.14	12.11 ± 13.01	0.00 ± 0.00
Causal CEM + CE	0	99.51 ± 0.05	5.67 ± 1.21	0.00 ± 0.00	12.41 ± 1.82	15.15 ± 4.10	0.00 ± 0.00

Table 8: Ablation study regarding λ_3 on the CelebA dataset.

	λ_3	Label Accuracy	Average CaCE	min CaCE	max CaCE	CaCE	CaCE block
Black Box		90.15 ± 1.30					
CBM		79.00 ± 0.18	0.54 ± 0.03	0.00 ± 0.00	1.67 ± 0.15	5.58 ± 2.36	5.46 ± 2.13
CEM		79.17 ± 0.26	0.27 ± 0.12	0.00 ± 0.00	1.07 ± 0.56	1.00 ± 0.45	1.06 ± 0.50
Causal CEM	0.05	78.23 ± 0.45	2.17 ± 1.44	0.00 ± 0.00	8.18 ± 4.38	0.04 ± 0.07	0.00 ± 0.00
Causal CEM + CE	0.05	78.42 ± 0.42	5.48 ± 0.34	0.00 ± 0.00	24.17 ± 1.32	1.24 ± 0.62	0.00 ± 0.00
Causal CEM	0.2	77.49 ± 0.37	1.70 ± 0.98	0.00 ± 0.00	8.95 ± 4.86	0.00 ± 0.00	0.00 ± 0.00
Causal CEM + CE	0.2	78.08 ± 0.39	6.15 ± 0.31	0.00 ± 0.00	29.57 ± 1.03	0.82 ± 0.46	0.00 ± 0.00
Causal CEM	0	77.42 ± 1.09	2.85 ± 0.44	0.00 ± 0.00	13.06 ± 2.65	0.00 ± 0.00	0.00 ± 0.00
Causal CEM + CE	0	78.31 ± 0.36	4.64 ± 0.13	0.00 ± 0.00	18.31 ± 2.61	1.25 ± 0.68	0.00 ± 0.00