entropy

MDPI

Article

Interpretability Is in the Mind of the Beholder: A Causal Framework for Human-Interpretable Representation Learning

Emanuele Marconato [1,2], Andrea Passerini [1] and Stefano Teso [1,3,*]

[1] Dipartimento di Ingegneria e Scienza dell'Informazione, University of Trento, 38123 Trento, Italy; emanuele.marconato@unitn.it (E.M.); andrea.passerini@unitn.it (A.P.)
[2] Dipartimento di Informatica, University of Pisa, 56126 Pisa, Italy
[3] Centro Interdipartimentale Mente/Cervello, University of Trento, 38123 Trento, Italy
* Correspondence: stefano.teso@unitn.it

Abstract: Research on Explainable Artificial Intelligence has recently started exploring the idea of producing explanations that, rather than being expressed in terms of low-level features, are encoded in terms of *interpretable concepts learned from data*. How to reliably acquire such concepts is, however, still fundamentally unclear. An agreed-upon notion of concept interpretability is missing, with the result that concepts used by both post hoc explainers and *concept-based* neural networks are acquired through a variety of mutually incompatible strategies. Critically, most of these neglect the human side of the problem: *a representation is understandable only insofar as it can be understood by the human at the receiving end*. The key challenge in human-interpretable representation learning (HRL) is how to model and operationalize this human element. In this work, we propose a mathematical framework for acquiring *interpretable representations* suitable for both post hoc explainers and concept-based neural networks. Our formalization of HRL builds on recent advances in causal representation learning and explicitly models a human stakeholder as an external observer. This allows us derive a principled notion of *alignment* between the machine's representation and the vocabulary of concepts understood by the human. In doing so, we link alignment and interpretability through a simple and intuitive *name transfer* game, and clarify the relationship between alignment and a well-known property of representations, namely *disentanglement*. We also show that alignment is linked to the issue of undesirable correlations among concepts, also known as *concept leakage*, and to content-style separation, all through a general information-theoretic reformulation of these properties. Our conceptualization aims to bridge the gap between the human and algorithmic sides of interpretability and establish a stepping stone for new research on human-interpretable representations.

Keywords: explainable AI; causal representation learning; alignment; disentanglement; causal abstractions; concept leakage

Citation: Marconato, E.; Passerini, A.; Teso, S. Interpretability Is in the Mind of the Beholder: A Causal Framework for Human-Interpretable Representation Learning. *Entropy* **2023**, *25*, 1574. https://doi.org/10.3390/e25121574

Academic Editors: Fabio Aiolli and Mirko Polato

Received: 10 September 2023
Revised: 31 October 2023
Accepted: 8 November 2023
Published: 22 November 2023

1. Introduction

The field of Explainable Artificial Intelligence (XAI) has developed a wealth of attribution techniques for unearthing the reasons behind the decisions of black-box machine learning models [1]. Traditionally, explaining a prediction involves identifying and presenting those low-level *atomic elements*—like input variables [2,3] and training examples [4,5]—that are responsible for said prediction (in the following, we will use the terms "responsibility" and "relevance" interchangeably). Explanations output by white-box models, such as sparse linear classifiers [6] and rule-based predictors [7], follow the same general setup. These atomic elements, however, are not very expressive and, as such, can be ambiguous [8]. To see this, consider an image of a red sports car that is tagged as "`positive`" by a black-box predictor. In this example, a saliency map would highlight those *pixels* that are most responsible for this prediction: these do not say whether the prediction depends on the image containing a "`car`", on the car being "`red`", or on the car being "`sporty`".

As a consequence, it is impossible to understand what the model is "thinking" and how it would behave on other images based on this explanation alone [9].

This is why focus in XAI has recently shifted toward explanations expressed in terms of higher-level symbolic representations, or *concepts* for short. These promise to ensure explanations are rich enough they can capture the machine's reasoning patterns, while being expressed in terms that can be naturally understood by stakeholders [8,10].

This trend initially emerged with (post hoc) *concept-based explainers* (CBEs) like TCAV [11] and Net2Vec [12], among others [13–15], which match the latent space of a deep neural network to a vocabulary of pre-trained concept detectors. The idea of using higher-level concepts was foreshadowed in the original LIME paper [3]. These were quickly followed by a variety of *concept-based models* (CBMs)—including Self-Explainable Neural Networks [16], Part-Prototype Networks [17], Concept-Bottleneck Models [18], GlanceNets [19], and Concept Embedding Models [20]—that support representation learning while retaining interpretability. Specifically, these approaches learn a neural mapping from inputs to concepts, and then leverage the latter for both computing predictions—in a simulatable manner [21]—and providing ante-hoc explanations thereof. See [22] for a review. Since concepts act as a *bottleneck* through which all information necessary for inference must flow, CBMs hold the promise of avoiding the lack of faithfulness typical of post hoc techniques, while enabling a number of useful operations such as interventions [18] and debugging [23,24] using concepts as a human-friendly interface.

1.1. Limitations of Existing Works

The premise of conceptual explanations rests on the assumption that learned concepts are themselves interpretable. This begs the question: *what does it mean for a vocabulary of concepts to be interpretable?*

Researchers have proposed a variety of practical strategies to encourage the interpretability of the learned concepts, but no consistent recipe. Some CBMs constrain their representations according to intuitive heuristics, such as similarity to concrete training examples [17] or activation sparsity [16]. However, the relationship between these properties and interpretability is unclear, and unsurprisingly, there are well-known cases in which CBMs acquire concepts activating on parts of the input with no obvious semantics [25,26]. A more direct way of controlling the semantics of learned concepts is to leverage *supervision* on the concepts themselves, a strategy employed by both CBEs [11] and CBMs [18,19,27]. Unfortunately, this is no panacea, as doing so cannot prevent *concept leakage* [28,29], whereby information from a concept "leaks" into another, seemingly unrelated concept, compromising its meaning.

At the same time, concept quality is either assessed qualitatively in a rather unsystematic fashion—e.g., by inspecting the concept activations or saliency maps on a handful of examples—or quantitatively, most often by measuring how well-learned concepts match annotations. This so-called *concept accuracy*, however, is insufficient to capture issues like concept leakage.

Besides these complications, existing approaches neglect a critical aspect of this learning problem: that *interpretability is inherently subjective*. For instance, explaining a prediction to a medical doctor requires different concepts than explaining it to a patient: the notion of "`intraepithelial`" may be essential for the former, while being complete gibberish to the latter. However, even when concept annotations are employed, they are gathered from offline repositories and as such they may not capture concepts that are meaningful to a particular expert, or that despite being associated with a familiar name, follow semantics incompatible with those the user attaches to that name. Of course, there are exceptions to this rule. These are discussed in Section 6.

1.2. Our Contributions

Motivated by these observations, *we propose to view interpretability as the machine's ability to communicate with a specific human-in-the-loop*. Specifically, we are concerned with the prob-

lem of learning conceptual representations that enable this kind of communication for both post and ante-hoc explanations. We call this problem **human-interpretable representation learning**, or HRL for short. Successful communication is essential for ensuring human stakeholders can understand *explanations* based on the learned concepts and, in turn, realizing the potential of CBEs and CBMs. This view is compatible with recent interpretations of the role of symbols in neuro-symbolic AI [10,30]. The key question is how to model this human element in a way that can be actually *operationalized*. We aim to fill this gap.

Our first contribution is a conceptual and mathematical model—resting on techniques from causal representation learning [31]—of HRL that *explicitly models the human-in-the-loop*.

As a second contribution, we leverage our formalization to develop an intuitive but sound notion of *alignment* between the conceptual representation used by the machine and that of the human observer. Alignment is strictly related to *disentanglement*, a property of learned representations frequently linked to interpretability [32,33], but also strictly *stronger*, in the sense that disentanglement alone is insufficient to ensure alignment. Later on, we will formally show that this follows from Proposition 1. We propose that alignment is key for evaluating interpretability of both CBEs and CBMs.

Our formalization improves on the work of Marconato et al. [19] and looks at three settings of increasing complexity and realism: (i) a simple but non-trivial setting in which the human's concepts are *disentangled* (i.e., individual concepts can be changed independently from each other without interference). (ii) a more general setting in which the human's concepts are constrained to be disentangled in blocks; and (iii) an unrestricted setting in which the human concepts can influence each other in arbitrary manners. In addition, we identify a and previously ignored link between interpretability of representations and the notion of *causal abstraction* [34–36].

As a third contribution, we formally show that *concept leakage* can be viewed as a lack of disentanglement, and therefore of alignment. This strengthens existing results and allows to reinterpret previous empirical observations [19,37].

As a fourth contribution, we discuss key questions arising from our mathematical framework, including whether perfect alignment is sufficient and necessary for interpretability, how to measure it, how to implement it in representation learning, and how to collect the necessary concept annotations.

1.3. Outline

The remainder of this paper is structured as follows. In the next section, we introduce prerequisite material, and then proceed in Section 3 to formalize the problem of human-interpretable representation learning and cast concept interpretability in terms of *alignment between representations*. Next, in Section 4, we analyze in depth the notion of alignment in three settings of increasing complexity and study its relationship to the issue of concept leakage, and then look at the consequences of our formalization in Section 5. Finally, we discuss related works in Section 6 and offer some concluding remarks in Section 7.

2. Preliminaries

In the following, we indicate scalar constants x in lower-case, random variables X in upper case, ordered sets of constants $\mathbf{x}$ and random variables $\mathbf{X}$ in bold typeface, and index sets $\mathcal{I}$ in calligraphic typeface. We also use the shorthand $[n] := \{1, \dots, n\}$. Letting $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathcal{I} \subseteq [n]$, we write $\mathbf{X}_\mathcal{I} := (X_i : i \in \mathcal{I})$ to indicate the ordered subset indexed by $\mathcal{I}$ and $\mathbf{X}_{-\mathcal{I}} := \mathbf{X} \setminus \mathbf{X}_\mathcal{I}$ to denote its complement, and abbreviate $\mathbf{X} \setminus \{X_i\}$ as $\mathbf{X}_{-i}$.

2.1. Structural Causal Models and Interventions

A *structural causal model* (SCM) is a formal description of the causal relationships existing between parts of a (stochastic) system [38,39]. Formally, an SCM $\mathfrak{C}$ specifies a set of *structural assignments* encoding direct causal relationships between variables (as customary, we work with SCMs that are *acyclic*, *causally sufficient* (i.e., there are no external,

hidden variables influencing the system), and *causally Markovian* (i.e., each variable X_i is independent of its non-descendant given its parents in the SCM) [38]) in the form:

$$X_i \leftarrow f_i(\mathbf{Pa}_i, N_i) \tag{1}$$

where $\mathbf{X} = (X_1, \ldots, X_n)$ are variables encoding the state of the system, $\mathbf{Pa}_i \subseteq \mathbf{X}$ are the direct causes of X_i, and N_i are noise terms. Variables without parents are *exogenous*, and play the role of inputs to the system, while the others are *endogenous*. The full state of the system can be sampled by propagating the values of the exogenous variables through the structural assignments in a top-down fashion. SCMs can be viewed as *graphs* in which nodes represent variables, arrows represent assignments, and noise variables are usually suppressed; see Figure 1.

Figure 1. SCMs illustrating two different notions of disentanglement. *Left*: the variables $\mathbf{G} = \{G_1, \ldots, G_n\}$ are disentangled. *Right*: Typical data generation and encoding process used in deep latent variable models. The machine representation $\mathbf{M} = \{M_1, \ldots, M_k\}$ is *disentangled with respect to* the generative factors $\mathbf{G}$ if and only if each M_j encodes information about one G_i at most.

Following common practice, we assume the noise terms to be mutually independent from each other and also independent from the variables not appearing in the corresponding structural equations; that is, it holds that $N_i \perp\!\!\!\perp N_j$ for all $i \neq j$ and $N_i \perp\!\!\!\perp X_j$ for all i, j. This is equivalent to assuming there are no hidden confounders. This assumption carries over to all SCMs used throughout the paper.

An SCM $\mathfrak{C}$ describes both a *joint distribution* $p(\mathbf{X}) = \prod_i p(X_i \mid \mathbf{Pa}_i)$ and how this distribution *changes* upon performing *interventions* on the system. These are modifications to the system's variables and connections performed by an external observer. Using Pearl's *do*-operator [38], (atomic) interventions can be written as $do(X_i \leftarrow x_i)$, meaning that the value of the variable X_i is forcibly changed to the value x_i, regardless of the state of its parents and children. Carrying out an atomic intervention yields a *manipulated SCM* identical to $\mathfrak{C}$, except that all assignments to X_i are deleted (i.e., the corresponding links in the graph disappear) and all occurrences of X_i in the resulting SCM are replaced by the constant x_i. The resulting manipulated distribution is $p(\mathbf{X} \mid do(X_i \leftarrow x_i)) = \mathbb{1}\{X_i = x_i\} \cdot \prod_{j \neq i} p(X_j \mid \mathbf{Pa}_j)$. Non-atomic interventions of the form $do(\mathbf{X}_\mathcal{I} \leftarrow \mathbf{x}_\mathcal{I})$ work similarly. Expectations of the form $\mathbb{E}[\cdot \mid do(X_j \leftarrow x_j)]$ are just regular expectations evaluated with respect to the manipulated distribution.

2.2. Disentanglement

Central to our work is the notion of disentanglement [31,33,40] in both its two acceptations, namely *disentanglement of variables* and *disentanglement of representations*. We henceforth rely on the causal formalization given by Suter et al. [41] and Reddy et al. [42]. We refer the reader to those papers for more details.

Intuitively, *a set of variables* $\mathbf{G} = (G_1, \ldots, G_n)$ *is disentangled if the variables can be changed independently from one another*. For instance, if G_1 represents the "`color`" of an object and G_2 its "`shape`", disentanglement of variables implies that changing the object's color does not impact its shape. This should hold even if the variables $\mathbf{G}$ have a common set of parents $\mathbf{C}$—playing the role of counfounders, such as sampling bias or choice of source domain [38]—meaning that they can be both disentangled *and* correlated (via $\mathbf{C}$). From a causal perspective, disentanglement of variables can be defined as follows:

Definition 1 (Disentanglement of variables)**.** *A set of variables* $\mathbf{G}$ *are disentangled if and only if* $p(G_i \mid \mathbf{C}, do(\mathbf{G}_{\mathcal{I}} \leftarrow \mathbf{g}'_{\mathcal{I}})) \equiv p(G_i \mid \mathbf{C})$ *for all possible choices of* $\mathcal{I} \subseteq [n] \setminus \{i\}$ *and* $\mathbf{g}'_{\mathcal{I}}$*.*

Now, consider the SCM in Figure 1 (left). It is easy to see that the variables $\mathbf{G}$ are disentangled: any intervention $do(\mathbf{G}_{\mathcal{I}} \leftarrow \mathbf{g}'_{\mathcal{I}})$ breaks the links from $\mathbf{C}$ to $\mathbf{G}_{\mathcal{I}}$, meaning that changes to the latter will not affect G_i. In this case, the variables $\mathbf{G}$ are also conditionally independent from one another given $\mathbf{C}$, or equivalently $G_i \perp\!\!\!\perp G_j \mid \mathbf{C}$ for every $i \neq j$.

Later on, we will be concerned with data generation processes similar to the one illustrated in Figure 1 (right). Here, a set of *generative factors* $\mathbf{G} = (G_1, \ldots, G_n)$ with common parents $\mathbf{C}$ cause an observation $\mathbf{X}$, and the latter is encoded into a *representation* $\mathbf{M} = (M_1, \ldots, M_k)$ by a machine learning model $p_\theta(\mathbf{M} \mid \mathbf{X})$. Specifically, the explicit relation between $\mathbf{M}$ and $\mathbf{G}$ is obtained by marginalizing over the inputs $\mathbf{X}$:

$$p_\theta(\mathbf{M} \mid \mathbf{G}) := \mathbb{E}_{\mathbf{x} \sim p(\mathbf{X} \mid \mathbf{G})}[p_\theta(\mathbf{M} \mid \mathbf{x})] \tag{2}$$

This can also be viewed as a *stochastic map* $\alpha : \mathbf{g} \mapsto \mathbf{m}$. Maps of this kind are central to our discussion.

Since $\mathbf{G}$ is disentangled (see Definition 1), we can talk about *disentanglement of representations* for $\mathbf{M}$. We say that $\mathbf{M}$ *is disentangled with respect to* $\mathbf{G}$ if, roughly speaking, each M_j encodes information about at most one G_i or, more precisely, *as long as* G_i *is kept fixed, the value of* M_j *does not change even when the remaining factors* $\mathbf{G} \setminus \{G_i\}$ *are forcibly modified via interventions*. The degree by which a representation *violates* disentanglement of representations can be measured using the PIDA metric:

Definition 2 (PIDA [41])**.** *Let* G_i *be a generative factor and* M_j *an element of the machine representation.* PIDA *measures how much fixing* G_i *to a given value* g_i *insulates* M_j *from changes to the other generative factors* $\mathbf{G}_{-i}$*, and it is defined as:*

$$\mathsf{PIDA}(G_i, M_j \mid g_i, \mathbf{g}_{-i}) := d\big(p_\theta(M_j \mid do(G_i \leftarrow g_i)), p_\theta(M_j \mid do(G_i \leftarrow g_i, \mathbf{G}_{-i} \leftarrow \mathbf{g}_{-i}))\big) \tag{3}$$

where d *is a divergence. The original definition [41] fixes* d *to be the difference between means. Here, we slightly generalize* PIDA *to arbitrary divergences, as doing so can account for changes in higher-order moments too. The average worst case over all possible choices of* g_i *and* $\mathbf{g}_{-i}$ *is given by:*

$$\mathsf{EMPIDA}(G_i, M_j) := \mathbb{E}_{g_i}[\max_{\mathbf{g}_{-i}} \mathsf{PIDA}(G_i, M_j \mid g_i, \mathbf{g}_{-i})] \tag{4}$$

Definition 3 (Disentanglement of representations)**.** *We say that a representation* $\mathbf{M}$ *is disentangled with respect to* $\mathbf{G}$ *if and only if* $\max_j \min_i \mathsf{EMPIDA}(G_i, M_j)$ *is exactly zero.*

In other words, $\mathbf{M}$ is disentangled with respect to $\mathbf{G}$ if, for every M_j there exists a G_i such that fixing the latter *insulates* M_j from changes to the other generative factors $\mathbf{G}_{-i}$. In Section 4, we will build on both types of disentanglement to derive our notion of alignment between representations.

Another important notion is that of *context-style separation*, which can be viewed as a special case of disentanglement of representations [43]. Let the generative factors $\mathbf{G}$ be partitioned into two disentangled sectors $\mathbf{G}_{\mathcal{I}}$ and $\mathbf{G}_{-\mathcal{I}}$, representing task-relevant information (content) and task-irrelevant factors of variations (style), respectively. Then, $\mathbf{M}$ satisfies content-style separation if the following holds:

Definition 4 (Content-style separation)**.** *Let* $(\mathbf{G}_{\mathcal{I}}, \mathbf{G}_{-\mathcal{I}})$ *be two disentangled sectors. Then,* $\mathbf{M}$ *separates content from style if it can be partitioned into* $(\mathbf{M}_{\mathcal{J}}, \mathbf{M}_{-\mathcal{J}})$ *such that:*

$$\mathsf{EMPIDA}(\mathbf{G}_{\mathcal{I}}, \mathbf{M}_{\mathcal{J}}) = 0 \tag{5}$$

This means that, if the content $\mathbf{G}_{\mathcal{I}}$ is fixed, the machine representation $\mathbf{M}_{\mathcal{J}}$ is isolated from changes to the style $\mathbf{G}_{-\mathcal{I}}$. This property is asymmetrical: it holds even if $\mathbf{M}_{-\mathcal{J}}$ *is* affected by interventions to $\mathbf{G}_{\mathcal{I}}$. Also, there is no requirement that the elements of $\mathbf{M}_{\mathcal{J}}$ are disentangled with respect to $\mathbf{G}_{\mathcal{I}}$.

3. Human Interpretable Representation Learning

We are concerned with acquiring interpretable machine representations. Our key intuition is that a representation is only interpretable as long as it can be *understood by the human at the receiving end*. Based on this, we formally state our learning problem as follows:

Definition 5 (Intuitive statement)**.** *Human-interpretable representation learning (*HRL*) is the problem of learning a (possibly stochastic) mapping between inputs* $\mathbf{x} \in \mathbb{R}^d$ *and a set of machine representations* $\mathbf{z} \in \mathbb{R}^k$ *that enables a machine and a specific human stakeholder to communicate using those representations.*

This mapping can be modeled without loss of generality as a conditional distribution $p_\theta(\mathbf{Z} \mid \mathbf{X})$, whose parameters θ are estimated from data. While Definition 5 encompasses both CBEs and CBMs, the meaning of $\mathbf{Z}$ differs in the two cases, as we show next.

3.1. Machine Representations: The Ante-Hoc Case

CBMs are neural predictors that follow the generative process shown in Figure 2 (left). During inference, a CBM observes an input $\mathbf{x}$, caused by generative factors $\mathbf{G}$, and extracts a representation $\mathbf{M}$, e.g., by performing *maximum a posteriori* inference [44], from the distribution $p_\theta(\mathbf{M} \mid \mathbf{x})$ implemented as a neural network. In practice, the concept encoder can be implemented in various ways, e.g., as a multi-label convolutional neural network [27], as a variational auto-encoder [19], or even as a large language model [45]. This representation is partitioned into two subsets: $\mathbf{M}_{\mathcal{J}}$ *are constrained to be interpretable, while* $\mathbf{M}_{-\mathcal{J}}$ *are not*. As shown in Figure 2, only the interpretable subset is used for inferring a prediction $\hat{y}$, while $\mathbf{M}_{-\mathcal{J}}$, if present, is used for other tasks, such as reconstruction [19]. Specifically, the predicted concepts $\mathbf{M}_{\mathcal{J}}$ are fed to a simulatable top layer $p_\theta(Y \mid \mathbf{M})$, most often a sparse linear layer, from which an explanation can be easily derived. Assuming $\mathbf{M}_{\mathcal{J}}$ is in fact interpretable, CBMs can provide local *explanations*, summarizing what concepts are responsible for a particular prediction in an ante-hoc fashion and essentially for free [22,46]. For instance, if $p_\theta(Y \mid \mathbf{M}_{\mathcal{J}})$ is a linear mapping with parameters w_{yj}, the explanation for predicting $\hat{y}$ is given by [16–19,27,47]:

$$\mathcal{E} = \{(w_{\hat{y}j}, m_j) \ : \ j \in \mathcal{J}\} \tag{6}$$

where each concept activation m_j is associated with a "level of responsibility" inferred from the top layer's weights. Specific CBMs are outlined in Section 6.

Summarizing, in the case of CBMs, the concepts $\mathbf{Z}$ used for communicating with users (see Definition 5) are embodied by the interpretable machine representation $\mathbf{M}_{\mathcal{J}}$.

3.2. Machine Representations: The Post Hoc Case

For CBEs, the generative process is different; see Figure 2 (right). In this case, the internal representation $\mathbf{M}$ of the model mapping from inputs $\mathbf{X}$ to labels Y *is not required to be interpretable*. For instance, it might represent the state of the neurons in a specific layer. CBEs explain the reasoning process in a post hoc fashion by extracting the activations of *high-level concepts* $\hat{\mathbf{H}}$ from $\mathbf{M}$, and then inferring a concept-based explanation $\mathcal{E}$ specifying the contribution of each $\hat{H}_i$ to the model's prediction, often in the same form as Equation (6).

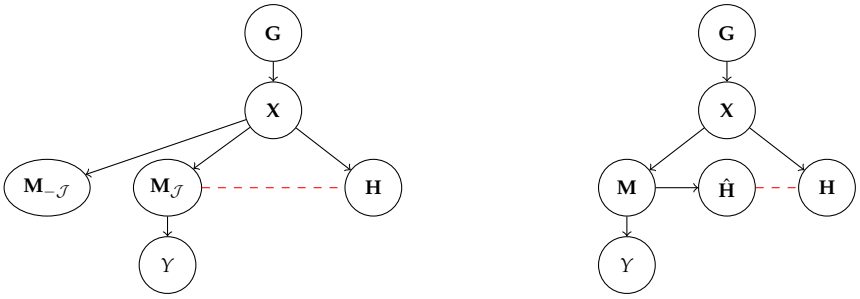

Figure 2. Left: following the generative process $p(\mathbf{X} \mid \mathbf{G})$, *concept-based models* (CBMs) extract a machine representation $\mathbf{M} = (\mathbf{M}_{\mathcal{J}}, \mathbf{M}_{-\mathcal{J}})$ via $p_\theta(\mathbf{M} \mid \mathbf{X})$, of which only $\mathbf{M}_{\mathcal{J}}$ is used to predict the label Y. $\mathbf{M}_{\mathcal{J}}$ contains all interpretable concepts, and as such it has to be *aligned* to their user concepts $\mathbf{H}$ (Section 4): the corresponding map is reported in red. **Right**: generative process followed by *concept-based explainers* (CBEs). Here, the machine representation $\mathbf{M}$ is *not* required to be interpretable. Rather, the concept-based explainer maps it to extracted concepts $\hat{\mathbf{H}}$ and then infers how these contribute to the prediction Y. Here, alignment should hold between $\hat{\mathbf{H}}$ and $\mathbf{H}$.

Here, we are concerned with the interpretability of $\hat{\mathbf{H}}$. Some approaches extract them by (indirectly) relying on concept annotations. For instance, TCAV [11] takes a set of linear classifiers, one for each concept, pre-trained on a densely annotated dataset, and then adapts them to work on machine representations $\mathbf{M}$. Unsupervised approaches instead mine the concepts directly in the space of machine representations through a linear decomposition [13–15,48]. Specific examples are discussed in Section 6. In general, there is no guarantee that the symbolic and sub-symbolic representations $\hat{\mathbf{H}}$ and $\mathbf{M}$ capture exactly the same information. This introduces a *faithfulness* issue, meaning that CBE explanations may not portray a reliable picture of the model's inference process [11,48–50].

However, the issue we focus on is that the representation $\mathbf{Z} = \hat{\mathbf{H}}$ used by CBEs to communicate with users is, in fact, interpretable, regardless of whether it is also faithful.

3.3. From Symbolic Communication to Alignment

What makes symbolic communication possible? While a complete answer to this question is beyond the scope of this paper, we argue that communication becomes challenging unless the concepts $\mathbf{Z}$ with which the machine and the human communicate are "*aligned*", in the sense that concepts having the same *name* share the same (or similar enough) *semantics*. Other factors contributing to interpretability, as well as some further remarks on whether alignment is sufficient and necessary, will be discussed in Section 5.

In order to formalize this intuition, we focus on the generative process shown in Figure 3. In short, we assume observations $\mathbf{x}$ (e.g., images or text observed during training and test) are obtained by mapping generative factors $\mathbf{G} \sim p^*(\mathbf{G} \mid \mathbf{C})$ through a hidden ground-truth distribution $p^*(\mathbf{X} \mid \mathbf{G})$.

The observations $\mathbf{x}$ are then received by *two observers*: a machine and a human. The machine maps them to its own learned representation $\mathbf{M}$, which may or may not be interpretable. The interpretable representations $\mathbf{Z}$, which correspond to $\mathbf{M}_{\mathcal{J}}$ for CBMs (see Section 3.1) and to $\hat{\mathbf{H}}$ for CBEs (Section 3.2), are then derived from $\mathbf{M}$.

At the same time, the human observer maps the same observations to its own vocabulary of concepts $\mathbf{H}$. For instance, if $\mathbf{x}$ is an image portraying a simple object on a black background, $\mathbf{h}$ may encode the "`color`" or "`shape`" of that object, or any other properties deemed relevant by the human. The choice and semantics of these concepts depend on the background and expertise of the human observer and possibly on the downstream task the human may be concerned with (e.g., medical diagnosis or loan approval) and, as such, may vary between subjects. It is *these* concepts that the human associates names, like in Figure 4, and it is these concepts that they would use for communicating the properties of $\mathbf{x}$ to other people.

Notice that the human concepts $\mathbf{H}$ may be arbitrarily different from the ground-truth factors $\mathbf{G}$: whereas the latter include all information necessary to determine the

observations, and as such may be complex and uninterpretable [51], the former are those aspects of the observation that matter *to the human observer*. A concrete example is that of color blindness: an observer may be unable to discriminate between certain wavelengths of visible light, despite these being causes of the generated image **X**. Another, more abstract, example is the generative factors that cause a particular apple to appear ripe, e.g., those biological processes occurring during the apple tree's reproductive cycle, which are beyond the understanding of most non-experts. They are so opaque that a whole science had to be developed to identify and describe them. In stark contrast, the concept of "`redness`" is not causally related to the apple's appearance, and yet easily understood by most human observers, precisely because it is a feature that is evolutionarily and culturally useful to those observers. In this sense, *the concepts* **H** *are understandable, by definition, to the human they belong to*.

We argue that symbolic communication is feasible whenever the names associated (by the human) to elements of **H** can be transferred to the elements of **Z** in a way that preserves semantics. That is, *concepts with the same name should have the same meaning*. In order to ensure information expressed in terms of **Z**—say, an explanation stating that Z_1 is irrelevant for a certain prediction—is understood by the human observer, we need to make sure that **Z** itself is somehow "aligned" with the human's representation **H**.

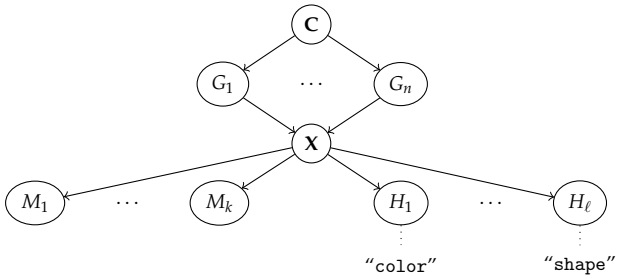

Figure 3. Graphical model of our data generation process. In words, n (correlated) generative factors exist in the world $\mathbf{G} = (G_1, \ldots, G_n)$ that *cause* an observed input **X**. The machine maps these to an internal representation $\mathbf{M} = (M_1, \ldots, M_k)$, while the human observer maps them to its own internal concept vocabulary $\mathbf{H} = (H_1, \ldots, H_\ell)$. Notice that the observer's concepts **H** may, and often do, differ from the ground-truth factors **G**. The concepts **H** are what the human can understand and attach names to, e.g., the "`color`" and "`shape`" of an object appearing in **X**. The association between names and human concepts is denoted by dotted lines. We postulate that communication is possible if the machine and the human representations are *aligned* according to Definition 6.

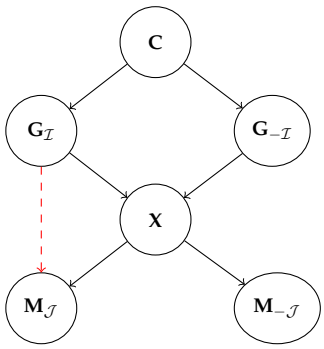

Figure 4. Simplified generative process with a single observer, adapted from [19]. Here, **C** is unobserved confounding variables influencing the generative factors **G**, and **M** is the latent representation learned by the machine. The red arrow represents the map α.

4. Alignment as Name Transfer

4.1. Alignment: The Disentangled Case

What does it mean for two representations to be aligned? We start by looking at the simplest (but non-trivial) case in which the ground-truth factors **G** are *disentangled*;

see Definition 1. For ease of exposition, let us also temporarily assume that some of the generative factors are inherently interpretable, as in [19]. Namely, we assume all factors in $\mathbf{G}_{\mathcal{I}} \subseteq \mathbf{G}$, where $\mathcal{I} \subseteq [n]$, can be understood by the human observer, while those in $\mathbf{G}_{-\mathcal{I}}$ cannot. The corresponding data generation process is illustrated in Figure 4. Under these assumptions, we aim to recover machine representations $\mathbf{M}$ that are aligned to the interpretable factors $\mathbf{G}_{\mathcal{I}}$.

To this end, we generalize the notion of alignment introduced by Marconato et al. [19]. Our definition extends that of [19] to the general case in which the mapping α, which is defined as a marginal distribution in Equation (2), is stochastic rather than deterministic. Doing so allows us to cater to more realistic applications and to draw an explicit connection with PIDA in Proposition 1. As anticipated, our definition revolves around the conditional distribution on $\mathbf{M}$ given by $\mathbf{G}$ or, equivalently, the stochastic map $\alpha : \mathbf{g} \mapsto \mathbf{m}$ defined in Equation (2) and shown in red in Figure 4. The key intuition is that *two concept vocabularies* $\mathbf{G}$ *and* $\mathbf{M}$ *are aligned if and only if* α *preserves the semantics of the interpretable generative factors* $\mathbf{G}_{\mathcal{I}}$.

More specifically, alignments holds if α allows to *transfer the names* of the interpretable factors in a way that preserves semantics. If $p_\theta(\mathbf{M} \mid \mathbf{X})$ is learned in an unsupervised fashion, names are generally transferred by collecting or constructing inputs annotated with the corresponding human concepts, feeding them to the concept extractor, and looking for matches between the annotations and the elements of $\mathbf{M}_{\mathcal{J}}$. If concept-level annotations are used, the names are automatically transferred along with them, but we still wish the user to be able to match the learned concepts with its own. In a sense, this process is analogous to giving the human observer access to a set of "knobs", each one controlling the value of one $G_i \in \mathbf{G}_{\mathcal{J}}$, and to a visualization of the machine representation $\mathbf{M}_{\mathcal{I}}$. Turning a knob is akin to *intervening* on the corresponding factor G_i. If, by turning a knob, the user is able to figure out what G_i corresponds to what M_j, then they will associate them with the same name. Since we are assuming $\mathbf{G}_{\mathcal{I}}$ is disentangled, turning one knob does not affect the others, which simplifies the process.

The formal definition of alignment is as follows:

Definition 6 (Alignment)**.** *Given generative factors* $\mathbf{G}$ *of which* $\mathbf{G}_{\mathcal{I}}$ *are interpretable, a machine representation* $\mathbf{M}$ *is aligned if the map* α *between* $\mathbf{G}$ *and* $\mathbf{M}$ *can be written as:*

$$\mathbf{M}_{\mathcal{J}} = \alpha(\mathbf{G}, \mathbf{N})_{\mathcal{J}} = (\mu_j(G_{\pi(j)}, N_j) \ : \ j \in \mathcal{J}) \tag{7}$$

where $\mathbf{M}_{\mathcal{J}} \subseteq \mathbf{M}$ *are the machine representations that ought to be interpretable,* $\mathbf{N}$ *are independent noise variables, and* π *and* μ *satisfy the following properties:*

D1. *The index map* $\pi : \mathcal{J} \mapsto \mathcal{I}$ *is surjective and, for all* $j \in \mathcal{J}$*, it holds that, as long as* $G_{\pi(j)}$ *is kept fixed,* M_j *remains unchanged even when the other generative factors* $\mathbf{G} \setminus \{G_{\pi(j)}\}$ *are forcibly modified.*

D2. *Each element-wise transformation* μ_j*, for* $j \in \mathcal{J}$*, is monotonic in expectation over* N_j*:*

$$\exists \bowtie \in \{>, <\} \text{ s. t. } \forall g'_{\pi(j)} > g_{\pi(j)}, \ \left(\mathbb{E}_{N_j}[\mu_j(g_{\pi(j)}, N_j)] - \mathbb{E}_{N_j}[\mu_j(g'_{\pi(j)}, N_j)]\right) \bowtie 0 \tag{8}$$

Let us motivate our two desiderata. In line with prior work on disentangled representations [32,33], **D1** requires that α should not "mix" multiple G_i's into a single M_j, regardless of whether the former belong to $\mathbf{G}_{\mathcal{I}}$ or not. For instance, if M_j blends together information about both color and shape, or about color and some uninterpretable factor, human observers would have trouble pinning down which one of their concepts it matches. If it does not, then turning the $G_{\pi(j)}$ knob only affects M_j, facilitating name transfer. The converse is not true: as we will see in Section 4.4, interpretable concepts with "compatible semantics" can in principle be blended together without compromising interpretability. We will show in Section 4.2 that this is equivalent to disentanglement.

D2 is also related to name transfer. Specifically, it aims to ensure that, whenever the user turns a knob $G_{\pi(j)}$, they can easily understand *what* happens to M_j and thus figure

out the two variables encode the same information. To build intuition, notice that both **D1** and **D2** hold for the *identity* function, as well as for those maps α that *reorder* or *rescale* the elements of $\mathbf{G}_\mathcal{I}$, which clearly preserve semantics and naturally support name transfer. Monotonicity captures all of these cases and also more expressive *non-linear* element-wise functions, while *conservatively* guaranteeing a human would be able to perform name transfer. Notice that the mapping needs not to be exact, in the sense that the output can depend on independent noise factors $\mathbf{N}$. This leaves room for stochasticity due to, e.g., variance in the concept learning step. Notice also that **D2** can be constrained further based on the application.

A couple of remarks are in order. First, of all, while we have defined alignment between interpretable factors $\mathbf{G}_\mathcal{I}$ and the machine representation $\mathbf{M}_\mathcal{J}$, the same definition works even if we replace the former with the user's concepts $\mathbf{H}$ and the latter with the concepts $\hat{\mathbf{H}}$ extracted by a concept-based explainer. In both cases, α is the map between human concepts $\mathbf{h}$ and either $\mathbf{m}_\mathcal{J}$ or $\hat{\mathbf{h}}$, and it is obtained by marginalizing over $\mathbf{X}$, $\mathbf{G}$, and $\mathbf{C}$ (see Figure 3). More generally, alignment can hold for *any* mapping between representations. We also observe that, since π maps $\mathcal{J}$ exclusively into $\mathcal{I}$, alignment entails a form of *content-style separation* (Definition 4), in that $\mathbf{M}_\mathcal{J}$ does not encode any information about $\mathbf{G}_{-\mathcal{I}}$. We will show in Section 4.3 that representations that do not satisfy this condition can be affected by *concept leakage*, while aligned representations cannot. Finally, we note that $\mathbf{M}$ can be aligned and still contain multiple transformations of the same $G_i \in \mathbf{G}_\mathcal{I}$. This does not compromise interpretability in that all "copies" can always be traced back to the same G_i. Practical considerations on how to measure *alignment* are discussed in Section 5.2.

4.2. Disentanglement Does Not Entail Alignment

Next, we clarify the relationship between alignment and disentanglement of representations by showing that the latter is exactly equivalent to **D1**:

Proposition 1. *Assuming noise terms are independent, as per Section 2, **D1** holds if and only if the representations are disentangled in* $(\mathbf{G}_\mathcal{I}, \mathbf{M}_\mathcal{J})$ *(see Definition 3).*

All proofs can be found in Appendix A. The equivalence between disentanglement of representations and **D1** implies that *disentanglement is insufficient for interpretability*: even if $\mathbf{M}$ is disentangled, i.e., each M_j encodes information about at most one $G_i \in \mathbf{G}_\mathcal{I}$, nothing prevents the transformation from G_i to its associated M_j from being arbitrarily complex, complicating name transfer. In the most extreme case, $\alpha(\cdot)_j$ may not be *injective*, making it impossible to distinguish between different g_is, or could be an arbitrary shuffling of the continuous line: this would clearly obfuscate any information present about G_i. This means that, during name transfer, a user would be unable to determine what value of M_j corresponds to what value of G_i or to anticipate how changes to the latter affect the former.

This is why **D2** in Definition 6 requires the map between each $G_i \in \mathbf{G}_\mathcal{I}$ and its associated M_j to be "simple". This extra desideratum makes alignment *strictly stronger* than disentanglement.

4.3. Alignment Entails No Concept Leakage

Concept leakage is a recently discovered phenomenon whereby the "interpretable" concepts $\mathbf{M}_\mathcal{J}$ unintentionally end up encoding information about extraneous concepts [29]. Empirically, leaky concepts are predictive for inference tasks that, in principle, do not depend on them. Situations like the following occur in practice, even if full concept supervision is used [19,28,29]:

Example 1. *Let* $\mathbf{X}$ *be a* `dSprites` *image [52] picturing a white sprite, determined by generative factors including "*`position`*", "*`shape`*", and "*`size`*", on a black background. Now imagine training a concept extractor* $p_\theta(\mathbf{M} \mid \mathbf{X})$ *so that* $\mathbf{M}_\mathcal{J}$ *encodes* `shape` *and* `size` *(but not* `position`*) by using full concept-level annotations for* `shape` *and* `size`*. The concept extractor is then frozen.*

During inference, the goal is to classify sprites as either positive ($Y = 1$) or negative ($Y = 0$) depending on whether they are closer to the top-right corner or the bottom-left corner. When concept leakage occurs, the label, which clearly depends only on `position`*, can be predicted with above random accuracy from* $\mathbf{M}_{\mathcal{J}}$*, meaning these concepts somehow encode information about* `position`*, which they are not supposed to.*

The issue with concept leakage is that it muddles the semantics of concepts $\mathbf{M}_{\mathcal{J}}$, which contain information they are not supposed to encode, and therefore of explanations built on them. Imagine that the learned concept for "`red`" also activates on a few objects that are, in fact, blue, due to leakage. Also, assume the predictor predicts a blue object as positive because `red` fires. Then, an explanation for that prediction would be that the blue object is positive because (according to the model) it is red. Clearly, this hinders trustworthiness. The only existing formal account of concept leakage was provided by Marconato et al. [19], who view it in terms of (lack of) out-of-distribution (OOD) generalization. Other works instead focus on in-distribution behavior and argue that concept leakage is due to encoding discrete generative factors using a continuous representation [37,53]. We go beyond these works by providing the first general formulation of concept leakage and showing that it is related to alignment. Specifically, we propose to view *concept leakage as a (lack of) content-style separation*, and show that this explains how concept leakage can arise both in- *and* out-of-distribution. In order to do so, we start by proposing a general reformulation of the concept leakage problem (Definition 7) and derive two bounds from mutual information properties (Proposition 2). Then, we show that a model that achieves perfect alignment avoids concept leakage entirely (Proposition 3 and Corollary 1).

We start by formalizing the intuition that concept leakage is excess prediction accuracy—gained by leveraging leaky concepts—compared to a leak-free baseline [19,53]. The corresponding generative process is reported in Figure 5. We assume the generative factors $\mathbf{G}$ are partitioned as $(\mathbf{G}_{\mathcal{I}}, \mathbf{G}_{-\mathcal{I}})$ such that *only* $\mathbf{G}_{-\mathcal{I}}$ are *informative* for predicting a label Y, mediated by the conditional distribution $p(Y \mid \mathbf{G}_{-\mathcal{I}})$. This implies that their mutual information is positive; that is, $I(\mathbf{G}_{-\mathcal{I}}, Y) > 0$. In this section, we are mostly concerned with the non-informativeness of $\mathbf{G}_{\mathcal{I}}$, hence we allow $\mathbf{G}_{-\mathcal{I}}$ to potentially contain also interpretable factors. Now, fix a concept encoder $p_\theta(\mathbf{M}_{\mathcal{J}} \mid \mathbf{X})$ and let $q_\lambda(Y \mid \mathbf{M}_{\mathcal{J}})$ be a predictor learned on top of it (in orange in the figure). *To quantify concept leakage, we look at how well the best possible such predictor can infer the label Y using* $\mathbf{M}_{\mathcal{J}}$ after intervening on $\mathbf{G}_{-\mathcal{I}}$. Analogously to EMPIDA (Definition 2), the intervention detaches $\mathbf{G}_{-\mathcal{I}}$ from $\mathbf{C}$, thus ensuring the label Y cannot be influenced by the irrelevant factors $\mathbf{G}_{\mathcal{I}}$. The resulting manipulated distribution on $\mathbf{G}$ is:

$$p'(\mathbf{G}) = p(\mathbf{G} \mid do(\mathbf{G}_{-\mathcal{I}} \leftarrow \mathbf{g}_{-\mathcal{I}}))q(\mathbf{g}_{-\mathcal{I}}) := \mathbb{E}_{\mathbf{C}}[p(\mathbf{G}_{\mathcal{I}} \mid \mathbf{C})]\mathbb{1}\{\mathbf{G}_{-\mathcal{I}} = \mathbf{g}_{-\mathcal{I}}\}q(\mathbf{g}_{-\mathcal{I}}) \tag{9}$$

where $q(\mathbf{g}_{-\mathcal{I}})$ is a distribution over possible interventions. This can be *any* distribution, with the only requirement that under any intervention $do(\mathbf{G}_{-\mathcal{I}} \leftarrow \mathbf{g}_{-\mathcal{I}})$, the model observes different variations in Y. If Y is constant, leakage is impossible, since $I(\mathbf{G}_{-\mathcal{I}}, Y) = 0$.

From the causal factorization in Figure 5, the joint probability of $(\mathbf{X}, Y)$ resulting from the post-interventional distribution $p'(\mathbf{G})$ is given by:

$$p(\mathbf{X}, Y) = \mathbb{E}_{\mathbf{g} \sim p'(\mathbf{G} \mid do(\mathbf{G}_{-\mathcal{I}} \leftarrow \mathbf{g}_{-\mathcal{I}}))}[p(Y \mid \mathbf{g}_{-\mathcal{I}})p(\mathbf{X} \mid \mathbf{g})] \tag{10}$$

Data of this kind appear, for example, in the `dSprites` experiment [19] outlined in Example 1. Here, during training, the "`position`" of the sprite is fixed (i.e., $\mathbf{G}_{pos} = \mathbf{G}_{-\mathcal{I}}$ are fixed to the center), while at test time the data contains different interventions over the position $\mathbf{G}_{pos} = \mathbf{G}_{-\mathcal{I}}$, and free variations in the other factors $\mathbf{G}_{\mathcal{I}}$ (e.g., "`shape`" and "`size`"). Essentially, these interventions move the sprite around the top-right and bottom-left borders, where the factors $\mathbf{G}_{pos}$ are extremely informative for the label Y.

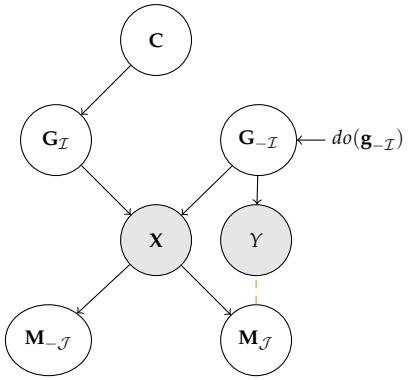

Figure 5. Generative process for Concept Leakage. A predictor observes examples $(\mathbf{X}, Y)$ and infers Y from its interpretable representation $\mathbf{M}_{\mathcal{J}}$ using a learnable conditional distribution $q_\lambda(Y \mid \mathbf{m}_{\mathcal{J}})$, indicated in orange. Since the label Y depends solely on $\mathbf{G}_{-\mathcal{I}}$, we would expect that it *cannot* be predicted better than at random: intuitively, if this occurs it means that information from $\mathbf{G}_{-\mathcal{I}}$ has leaked into the interpretable concepts $\mathbf{M}_{\mathcal{J}}$. Any intervention $do(\mathbf{G}_{-\mathcal{I}} \leftarrow \mathbf{g}_{-\mathcal{I}})$ on the uninterpretable/unobserved concepts detaches these from $\mathbf{C}$, meaning that the label truly only depends on $\mathbf{G}_{-\mathcal{I}}$.

In order to measure the degree of concept leakage in $p_\theta(\mathbf{M}_{\mathcal{J}} \mid \mathbf{X})$, *we compare the prediction performance of the best possible predictor* $q_\lambda(Y \mid \mathbf{M}_{\mathcal{J}})$ *with that of the best possible predictor* $r_\gamma(Y)$ *that does not depend on* $\mathbf{M}_{\mathcal{J}}$ *at all*. This is equivalent to comparing the behavior of two Bayes optimal predictors, one of which has access to the learned (possibly leaky) concepts whereas the other does not. In the following, we assume the distributions q_λ an r_γ to be sufficiently expressive, i.e., they can encode any sufficiently well behaved stochastic function. This is the case, for instance, when they are implemented as deep neural networks. We are now ready to define concept leakage:

Definition 7 (Concept leakage)**.** *Given a classifier* $q_\lambda(y \mid \mathbf{z})$*, an uninformed Bayes optimal predictor* $r_\gamma(y)$*, and data samples* $(\mathbf{x}, y) \in \mathcal{D}$*, concept leakage* Λ *is the difference:*

$$\Lambda = \max_{\lambda}[\mathcal{L}_{CL}(\lambda)] - \max_{\gamma}[\mathcal{L}_r(\gamma)] \tag{11}$$

where:

$$\mathcal{L}_{CL} = \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{X},Y)} \log q_{\lambda,\theta}(y|\mathbf{x}) \qquad \mathcal{L}_r = \mathbb{E}_{(\mathbf{x},y)\sim p(\mathbf{X},Y)} \log r_\gamma(y) \tag{12}$$

are the average log-likelihood of the classifier $q_{\lambda,\theta}(Y|\mathbf{X}) := \mathbb{E}_{\mathbf{m}_{\mathcal{J}}\sim p_\theta(\mathbf{M}_{\mathcal{J}}|\mathbf{X})} p(Y \mid \mathbf{m}_{\mathcal{J}})$ *and of the uninformed Bayes optimal classifier, respectively.*

By Definition 7, concept leakage occurs if and only if there exists a λ that allows to predict Y better than the best uninformed predictor. In the following analysis, we characterize concept leakage evaluated on the ground-truth distribution $p(\mathbf{X}, Y)$. We proceed to show that this quantity is bounded by two terms:

Proposition 2. *Assuming the causal factorization in Figure 5, it holds that:*

$$I(\mathbf{M}_{\mathcal{J}}, Y) \leq \Lambda \leq I(\mathbf{G}_{-\mathcal{I}}, Y) \tag{13}$$

where $I(\mathbf{A}, \mathbf{B})$ *denotes the mutual information between* $\mathbf{A}$ *and* $\mathbf{B}$*.*

The bounds in Equation (13) are useful for understanding how concept leakage behaves. They show, for instance, that Λ cannot exceed the mutual information between $\mathbf{G}_{-\mathcal{I}}$ and Y. Second, applying the data-processing inequality [54] to the lower bound yields $I(\mathbf{M}_{\mathcal{J}}, Y) \geq I(\mathbf{M}_{\mathcal{J}}, \mathbf{G}_{-\mathcal{I}})$. The latter quantifies the information contained in $\mathbf{M}_{\mathcal{J}}$ about $\mathbf{G}_{-\mathcal{I}}$. In other words, concept leakage can only be zero if indeed the machine concepts

$\mathbf{M}_{\mathcal{J}}$ contain no information about them, because $I(\mathbf{M}_{\mathcal{J}}, \mathbf{G}_{-\mathcal{I}}) \leq \Lambda = 0$. Next, we also show that if $\mathbf{M}_{\mathcal{J}}$ does not encode information about $\mathbf{G}_{-\mathcal{I}}$—or equivalently, it satisfies content-style separation (Definition 4)—then it has zero concept leakage.

Proposition 3. *Suppose that $\mathbf{M}_{\mathcal{J}}$ does not encode any information of $\mathbf{G}_{-\mathcal{I}}$, consistently with content-style separation (Definition 4), then Λ is zero.*

This result leads to two consequences. Let us start by looking at the *out-of-distribution case* investigated in [19]. Here, the concept extractor is trained only on some fixed variations of $\mathbf{G}_{-\mathcal{I}}$. However, when the support of $\mathbf{G}_{-\mathcal{I}}$ changes drastically, the model is not likely to ensure content-style separation outside of the support of the training distribution, even if **D1** holds in-distribution. This failure can be explained by the difficulty of disentanglement techniques to ensure disentanglement for out-of-distribution samples, in the context of combinatorial generalization, by Montero et al. [55,56]. Consider the dSprites example. Here, during training, sprites are located in the dead center of the background, and when observing the sprites on the borders of the image, which is far away from the support of the training set, the concept encoder fails to ensure their representations are disentangled. This failure of disentanglement techniques to ensure disentanglement for out-of-distribution inputs was also observed, in the context of combinatorial generalization, by Montero et al. [55,56]. Our results show that if content-style separation does not hold, concept leakage may be non-zero, meaning that techniques like open-set recognition [57] must be adopted to detect OOD inputs and process them separately.

Next, we look at concept leakage for *in-distribution* scenarios. Following Havasi et al. [53], consider a model leveraging two concepts—the presence of "`tail`" and "`fur`"—and the task of distinguish between images of cats and dogs using these (clearly non-discriminative) concepts. According to [53], concept leakage can occur when binary concepts like these are modeled using continuous variables, meaning the concept extractor can unintentionally encode "spurious" discriminative information. In light of our analysis, we argue that concept leakage is instead due to lack of content-style separation, and thus of alignment. To see this, suppose there exists a concept $G_k \in \mathbf{G}_{-\mathcal{I}}$ useful for distinguishing cats from dogs and that it is disentangled as in Definition 1 from the concepts of fur G_{fur} and of tail G_{tail}. Then, by content-style separation, any representation $\mathbf{M}_{\mathcal{J}}$ that is aligned to G_{fur} and G_{tail} does not encode any information about G_k, leading to zero concept leakage.

In both cases, concept leakage arises as a failure in content-style separation between relevant and irrelevant generative factors, and as such it can be used as a proxy for measuring the latter. Moreover, since alignment implies content-style separation, aligned representations cannot suffer from concept leakage. Note that the converse is not true: while alignment entails content-style separation, the latter can hold independently from alignment.

In-distribution concept leakage can also be extended to encompass the case where concepts belong to interpretable concepts $\mathbf{G}_{\mathcal{I}}$. This is the original context considered by Havasi et al. [53] where, for instance, $(G_{fur}, G_{tail}, G_k) \subseteq \mathbf{G}_{\mathcal{I}}$. Again, we suppose that only the ground-truth factor G_k is relevant for in-distribution predictions Y (being a cat or a dog). In this case, concept leakage is evaluated among those elements of $\mathbf{M}_{\mathcal{J}}$ that should not encode G_k. Practically, if a subset of $\mathbf{M}_{\mathcal{J}}$ encodes only the concepts G_{fur} and G_{tail} it must not be discriminative for Y, otherwise complicating the semantics of the learned concepts. Without loss of generality (the general case includes all representations $\mathbf{M}_{\pi^{-1}(k)}$, where π^{-1} is the pre-image of the map π), we suppose that only a single $M_{j'}$ is aligned to G_k, that is $\pi(j') = k$, whereas other $\mathbf{M}_{\mathcal{J}} \setminus M_{j'}$ are aligned to other concepts, among which G_{fur} and G_{tail}. Then, the following holds:

Corollary 1. *Consider a representation $\mathbf{M}_{\mathcal{J}}$ that is aligned to a set of disentangled concepts $\mathbf{G}_{\mathcal{I}}$, among which only G_k is discriminative for the label Y. Then, all $M_j \in \mathbf{M}_{\mathcal{J}}$ that are not associated by α to G_k, i.e., $\pi(j) \neq k$, do not suffer from concept leakage.*

Ultimately, an aligned representation prevents concept leakage within the encoded concepts $\mathbf{M}_{\mathcal{J}}$, guaranteed by having a *disentanglement* from **D1**. In fact, if the representations (M_{j_1}, M_{j_2}) are aligned to the concepts G_{fur} and G_{tail}, respectively, they cannot be used to discriminate between *cats* and *dogs*.

4.4. Alignment: The Block-Wise Case

So far, we assumed the generative factors $\mathbf{G}_{\mathcal{I}}$ (or, equivalently, the human concepts $\mathbf{H}$) are disentangled. We now extend alignment to more complex cases in which the human concepts *can* be mixed together without compromising interpretability. This covers situations in which, for instance, the machine captures a single categorical generative factor using multiple variables via one-hot encoding, or uses polar coordinates to represent the 2D position of an object.

To formalize this setting, we assume $\mathbf{G}_{\mathcal{I}}$ and $\mathbf{M}_{\mathcal{J}}$ are partitioned into non-overlapping "blocks" of variables $\mathbf{G}_{\mathcal{I}'} \subseteq \mathbf{G}_{\mathcal{I}}$ and $\mathbf{M}_{\mathcal{J}'} \subseteq \mathbf{M}_{\mathcal{J}}$, respectively. The idea is that each block $\mathbf{M}_{\mathcal{J}'}$ captures information about only a single block $\mathbf{G}_{\mathcal{I}'}$, and that while mixing across blocks is not allowed, mixing the variables within each block *is*. From the human's perspective, this means that name transfer is now performed block by block. With this in mind, we define *block alignment* as follows:

Definition 8 (Block-wise alignment)**.** *A machine representation* $\mathbf{M}$ *is block-wise aligned to* $\mathbf{G}_{\mathcal{I}}$ *if and only if there exists a subset* $\mathbf{M}_{\mathcal{J}} \subseteq \mathbf{M}$*, a partition* $\mathcal{P}_{\mathbf{M}}$ *of* $\mathcal{J}$*, and a mapping* $\alpha : (\mathbf{g}, \mathbf{N}) \mapsto \mathbf{m}$ *such that:*

$$\mathbf{M}_{\mathcal{J}'} = \alpha(\mathbf{G}, \mathbf{N})_{\mathcal{J}'} := \mu_{\mathcal{J}'}(\mathbf{G}_{\Pi(\mathcal{J}')}, \mathbf{N}_{\mathcal{J}'}) \qquad \forall \mathcal{J}' \in \mathcal{P}_{\mathbf{M}} \tag{14}$$

where the maps Π *and* μ *satisfy the following properties.*

- ***D1*** *There exists a partition* $\mathcal{P}_{\mathbf{G}}$ *of* $\mathcal{I}$ *such that* $\Pi : \mathcal{P}_{\mathbf{M}} \to \mathcal{P}_{\mathbf{G}}$*. In principle, we can extend this notion to a family of subsets* $\mathcal{P}_{\mathbf{G}}$ *of* $\mathcal{I}$*. As an example, for* xyz *positions, one can consider blocks* $\{xy, yz, xz\}$ *that are mapped to, respectively, block aligned representations. We call this condition block-wise disentanglement.*
- ***D2*** *Each map* $\mu_{\mathcal{J}'}$ *is simulatable and invertible (for continuous variables, we require it to be a diffeomorphism) on the first statistical moment; that is, there exists a unique pre-image* α^{-1} *defined as:*

$$\mathbf{G}_{\Pi(\mathcal{J}')} = \alpha^{-1}(\mathbb{E}[\mathbf{M}_{\mathcal{J}}])_{\mathcal{J}'} := \left(\mathbb{E}_{\mathbf{N}_{\mathcal{J}}}[\mu_{\mathcal{J}'}(\cdot, \mathbf{N}_{\mathcal{J}'})]\right)^{-1}(\mathbf{G}_{\Pi(\mathcal{J}')}) \tag{15}$$

By **D1**, changes to any block of human concepts only impact a single block of machine concepts, and by **D2** the change can be anticipated by the human observer, that is the human interacting with the machine grasps what is the general mechanism behind the transformation from $\mathbf{G}$ (or $\mathbf{H}$) and $\mathbf{M}$ (and vice versa). Both properties support name transfer.

A priori, it is not clear to say what transformations are simulatable [21], as this property depends crucially on the human's cognitive limitations and knowledge. We remark that **D2** intuitively constraints the variables within each block to be "semantically compatible". In the context of image recognition, for instance, placing concepts such as "`nose shape`" and "`sky color`" in the same block is likely to make name transfer substantially more complicated, as changes to "`nose shape`" might end affecting the representation of "'`sky color`" and vice versa. In that case, it would not be easy for a user to figure out how the concepts have been mixed, undermining simulatability. An example of semantic compatibility is that of rototranslation of the coordinates followed by element-wise rescaling. This condition is identical to "weak identifiability" in representation learning [58,59]. A counter example would be a map α given by a conformal map for the 2D position of an object in a scene. Albeit invertible, it may not be simple at all to simulate.

Notice that with this definition, we include two possible scenarios: (i) the case where some of the ground-truth concepts belonging to the same block are transformed in a single block; and (ii) the case where semantically compatible, but disentangled, concepts $\mathbf{G}_I$

are mixed together in $\mathbf{M}_{\mathcal{J}}$, which is often neglected in current disentanglement literature. The latter includes and extends the special case of alignment for disentangled $\mathbf{G}$.

The limitation of Definition 8 reflects the fact that taking into account the possible user's grasp of the representation is not straightforward to define and poses a challenge to provide a uniquely accepted definition that considers the human factor.

4.5. Alignment: The General Case

In the most general case, the generative factors $\mathbf{G}$ are causally related to each other according to an arbitrary ground-truth SCM $\mathfrak{C}_{\mathbf{G}}$. This entails that $\mathbf{G}_{\mathcal{I}}$ is no longer (block) disentangled. Hence, during name transfer, turning a "knob" (i.e., a variable G_i) affects those knobs that are causally dependent on it.

Naturally, the semantics of the user's comprises the causal relations between them. To see this, let G_1 be the `temperature` and G_2 the `color` of a metal object: the user knows that temperature affects color, and would not assign the same name to a representation of temperature that does not have a similar effect on the representation of color. In order ensure preservation of these semantics, we say that a machine representation $\mathbf{M}$ is *aligned* to $\mathbf{G}$ if, whenever the human intervenes on one G_i, affecting those generative factors $\mathbf{G}_{\mathcal{I}'}$ that depend on it, an analogous change occurs in the machine representation.

We now show that *block-alignment* is sufficient to satisfy this desideratum. Even in this more general case, the distribution over the representation can be obtained by marginalizing over the input variables:

$$p(\mathbf{M}) := \mathbb{E}_{\mathbf{x}\sim p(\mathbf{X})} p_\theta(\mathbf{M} \mid \mathbf{x}) \equiv \mathbb{E}_{\mathbf{g}\sim p(\mathbf{G})} p_\theta(\mathbf{M} \mid \mathbf{g}) \tag{16}$$

Notice that the definition of block alignment does not make any assumption about absence or presence of causal relations between blocks of generative factors, meaning that we it is still well-defined in this more general setting. This generalizes block alignment beyond disentangled factors [41]. Specifically, a map α can be block aligned if the variables G within each block are disentangled with each other, although there may exist causal relations across blocks.

Now, imagine having a stochastic map α between $\mathbf{G}$ and $\mathbf{M}$ that does satisfy block alignment, and also that there exist causal relations between the blocks $\mathbf{G}_{\mathcal{I}'}$. Whenever the user turns a "knob" corresponding to a ground-truth block, this yields an interventional distribution $p(\mathbf{G} \mid do(\mathbf{G}_{\mathcal{I}'} \leftarrow \mathbf{g}_{\mathcal{I}'}))$. Through α, this determines a new interventional distribution on the machine representations, namely:

$$p(\mathbf{M} \mid do(\mathbf{G}_{\mathcal{I}'} \leftarrow \mathbf{g}_{\mathcal{I}'})) = \mathbb{E}_{\mathbf{g}\sim p(\mathbf{G}\mid do(\mathbf{G}_{\mathcal{I}'}\leftarrow \mathbf{g}_{\mathcal{I}'}))} p_\theta(\mathbf{M} \mid \mathbf{g}) \tag{17}$$

This implies a representation $\mathbf{M}$ where the (interventional) distribution is obtained by mapping the state $\mathbf{g}_{\mathcal{I}'}$ through α. The same operation can be performed to obtain the state of all other machine representations aligned with the blocks that are causally related to $\mathbf{G}_{\mathcal{I}'}$ and affected by the intervention.

Note that this distribution *automatically takes causal relations between generative factors into account and treats them as causal relations between machine representations*. To see this, consider the following example:

Example 2. *Consider two generative factors G_1 and G_2 causally connected via a structural assignment $G_2 \leftarrow f(G_1, N_2)$, as in Figure 6. As before, G_1 could be the* `temperature` *and G_2 the* `color` *of a metal solid. Correspondingly, the aligned representation $\mathbf{M}$ encodes the* `temperature` *in two distinct variables, M_1 and M_3, respectively, to the temperature, say, measured in degrees Celsius and degrees Fahrenheit. M_2 encodes the* `color` *variable.*

The consequence of block-wise alignment is sketched in Figure 6 in three distinct cases: (left) intervening on the temperature G_1 affects both the aligned variables (M_1, M_3) and the color G_2. Correspondingly, this also has an effect on M_2 that changes according to G_2. (center) An intervention on G_2 influences only M_2 through α and it does not affect M_1 and M_3. (right) The

effect of an intervention on whole variables **G** *is localized such that interventions on the temperature factor* G_1 *will affect* M_1 *and* M_3*, whereby the interventions on* G_2 *only affect* M_2*, isolating it from the intervention on* G_1*.*

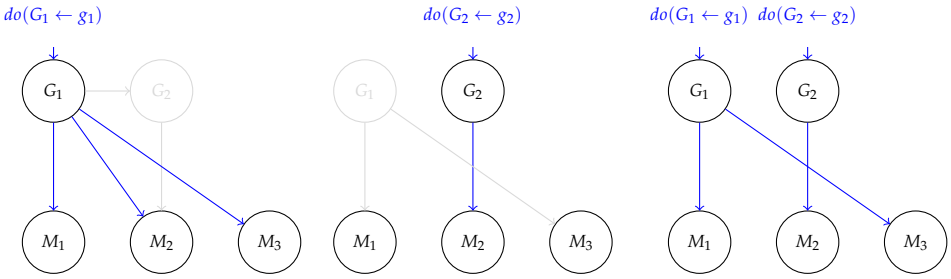

Figure 6. Block-aligned representation when $\mathfrak{C}_{\mathbf{G}}$ has causal connections. (**left**) An intervention on G_1 affects all representation (displayed in blue), since (M_1, M_3) are block-aligned to G_1 and M_2 is aligned to G_2. (**center**) Conversely, an intervention on G_2 only affects M_2, leaving the remaining representations untouched. (**right**) Intervening on all **G** has the effect of isolating the corresponding aligned representations from other interventions. In this case, intervening on G_2 removes the causal connection with G_1, so that M_2 does not depend on the intervention of G_1. Refer to Example 2 for further details.

Next, we formalize our observation that, thanks to *block alignment*, interventions on **G** are automatically mirrored on **M**:

Proposition 4. *Given a block-wise aligned representation* **M** *to* **G***, it holds that for each distinct block* $\mathbf{M}_{\mathcal{K}}$ *of representations* **M***, an intervention on* $\mathbf{G}_{\Pi(\mathcal{K})}$ *isolates* $\mathbf{M}_{\mathcal{K}}$ *from interventions of other ground-truth blocks* $\mathbf{G}_{-\Pi(\mathcal{K})}$*. Moreover, distinct interventions on* $\mathbf{G}_{\Pi(\mathcal{K})}$ *corresponds on average to different interventions on* $\mathbf{M}_{\mathcal{K}}$*.*

Importantly, this means that the effect of an intervention on the whole **G** isolates each block in **M** from the others, i.e., there is no explicit causal relation appearing in the learned representation. This matches the intuition that an intervention on a specific generative factor affects the corresponding block and removes the dependencies on other blocks of the representation. Following Example 2, this means that the a intervention on the `temperature` concept will affect both the corresponded representations and the ones aligned to concept of `color`, whereas intervening on the latter would only amount to change the representation of `color`, irrespective of the value assumed in the `temperature` representation.

To summarize, block alignment entails interventions on the ground-truth concepts are mapped properly. At the same time, alignment between blocks ensures the transformation α is simulatable, meaning that users can understand changes happening to all of the variables involved. This is sufficient to guarantee name transfer can be completed successfully in the general case, assuming not too many factors are changed at a time.

4.6. Alignment and Causal Abstractions

One important observation is that the form of name transfer we have considered is *asymmetrical*, in the sense that the user intervenes on its own representation **H** only, to then check how this impacts **M**. The other direction is not considered: it is not necessary to consider how intervening on **M** impacts **M**. This leads to the setup depicted in Figure 7 (right) in which, given $\mathfrak{C}_{\mathbf{H}}$, the effects of interventions on H_i are propagated to **M** via a map $\beta : \mathbf{H} \mapsto \mathbf{M}$, which may or may not be block-aligned.

We now consider a scenario in which the SCM of the representation $\mathfrak{C}_{\mathbf{M}}$ is also provided (in practice, $\mathfrak{C}_{\mathbf{M}}$ can be uncovered from data via causal discovery [38]) and the effects of interventions on **M** can be propagated leveraging its structural assignments.

Ideally, we would expect that, as long as $\mathbf{M}$ is block-aligned to $\mathbf{H}$, we can always find analogous post-interventional effects when intervening on H_i and on its aligned variable M_j. This underlies a consistency condition between the two "worlds" that are described with $\mathfrak{C}_\mathbf{H}$ and $\mathfrak{C}_\mathbf{M}$, respectively, by requiring that they both lead to similar conclusions when intervened in a equivalent manner. Clearly, this does not depend solely on the nature of the map β, but also on the structure of the machine SCM $\mathfrak{C}_\mathbf{M}$.

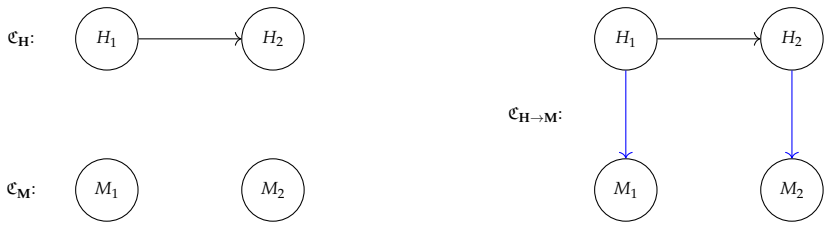

Figure 7. Absence of aligned causal abstraction. (**left**) The user's $\mathfrak{C}_\mathbf{H}$ incorporates a causal connection between H_1 to H_2, while the machine one $\mathfrak{C}_\mathbf{M}$ presents no causal connections. (**right**) The total SCM $\mathfrak{C}_{\mathbf{H}\to\mathbf{M}}$ of user's and machine's concepts resulting from an **aligned** map $\beta : \mathbf{H} \to \mathbf{M}$ (in blue). Refer to Example 3 for further discussion.

The presence of a consistency property between $\mathfrak{C}_\mathbf{H}$ and $\mathfrak{C}_\mathbf{M}$ is what defines a *causal abstraction* [34,35,60]; see [61] for an overview. Causal abstractions have been proposed to define (approximate) equivalence between causal graphs and have recently been employed in the context of explainable AI [36,62]. The existence of a causal abstraction ensures two systems are *interventionally equivariant*: interventions on one system can always be mapped (modulo approximations) to equivalent interventions in the other and lead to the same interventional distribution.

All causal abstractions check the consistency between two maps under the same intervention $do(\mathbf{H}_\mathcal{I})$: one is defined by the post-interventional distribution of $\mathfrak{C}_\mathbf{H}$ that is mapped on $\mathbf{M}$ via β, the other one consists of first matching on $\mathbf{M}$ the correspondent action $do(\mathbf{M}_\mathcal{J})$ and propagating it via $\mathfrak{C}_\mathbf{M}$. Intuitively, this means that, under β, interventions on $\mathbf{H}$ lead to the same conclusion as interventions on $\mathbf{M}$. We formalize this idea from constructive causal abstractions (existing works on causal abstractions [34,62] do not impose the map between values or interventions are simulatable, meaning that even if $\mathfrak{C}_\mathbf{M}$ is a causal abstraction of $\mathfrak{C}_\mathbf{H}$, it may be impossible for users to understand the mapping between the two) [62] by adapting it to the case where $\mathbf{H}$ and $\mathbf{M}$ are connected by block-alignment:

Definition 9 (β-aligned causal abstraction)**.** *The $\mathfrak{C}_\mathbf{M}$ is a causal abstraction of $\mathfrak{C}_\mathbf{H}$ under block-alignment β if, for all possible interventions $do(\mathbf{H}_\mathcal{I} \leftarrow \mathbf{h}_\mathcal{I})$ with $\mathbf{H}_\mathcal{I} \subseteq \mathbf{H}$, the following diagram commutes:*

$$\begin{array}{ccc} do(\mathbf{H}_\mathcal{I} \leftarrow \mathbf{h}_\mathcal{I}) & \xrightarrow{\mathfrak{C}_\mathbf{H}} & p(\mathbf{H} \mid do(\mathbf{H}_\mathcal{I} \leftarrow \mathbf{h}_\mathcal{I})) \\ \downarrow \beta & & \downarrow \beta_* \\ do(\mathbf{M}_\mathcal{J} \leftarrow \mathbf{m}_\mathcal{J}) & \xrightarrow{\mathfrak{C}_\mathbf{M}} & p(\mathbf{M} \mid do(\mathbf{M}_\mathcal{J} \leftarrow \mathbf{m}_\mathcal{J})) \end{array} \tag{18}$$

where β_* denotes the push-forward operation applied to the probability $p(\mathbf{H} \mid do(\mathbf{H}_\mathcal{I} \leftarrow \mathbf{h}_\mathcal{I}))$, and $\mathcal{J} = \Pi^{-1}(\mathcal{I})$ is the pre-image of $\mathcal{I}$ under Π.

In other words, aligned causal abstractions extend block alignment by enforcing a *symmetrical* consistency condition over interventions when both SCMs $\mathfrak{C}_\mathbf{H}$ and $\mathfrak{C}_\mathbf{M}$ are known: interventions on $\mathbf{M}$ have analogues on $\mathbf{H}$ and vice versa, i.e., Equation (18) holds and is *commutative*. This becomes relevant in situations where the user cannot parse the effect of an intervention on H_i on the input $\mathbf{X}$, i.e., they do not have access to $p(\mathbf{X} \mid do(H_i \leftarrow h_i))$, and they are left to validate the effects of their actions through β. In this case, leveraging on the SCM $\mathfrak{C}_\mathbf{M}$, the user can check how the mirrored intervention on M_j spreads

in the machine representations, and compare it with the corresponding representations given by β when the intervention is propagated on the user's factors $\mathbf{H}_{-i}$.

Therefore, while a map β being aligned is a necessary condition, it is not sufficient to guarantee a successful *name transfer* if $\mathfrak{C}_{\mathbf{M}}$ is highly dissimilar from $\mathfrak{C}_{\mathbf{H}}$. We show this situation explicitly where, despite having alignment between the user and the machine, the consistency condition in Equation (18) does not hold.

Example 3. *We consider two SCMs, one over user variables* $\mathfrak{C}_{\mathbf{H}}$ *and one over the machine ones* $\mathfrak{C}_{\mathbf{M}}$*. As shown in Figure 7 (left), the two SCMs have a different structure and for ease of reference we refer to* H_1 *and* M_1 *as the* `temperature` *variable and to* H_2 *and* M_2 *as the* `color` *variable. Despite the different structure, we suppose* M_1 *and* M_2 *are aligned to* H_1 *and* H_2*, respectively, via an aligned map* β*. We indicate the overall causal graph as* $\mathfrak{C}_{\mathbf{H}\to\mathbf{M}}$*; see Figure 7 (right).*

We can now check that $\mathfrak{C}_{\mathbf{M}}$ *is not an aligned abstraction of* $\mathfrak{C}_{\mathbf{H}}$ *under* β*. In fact, intervening on* H_1 *leads to different results on* $\mathfrak{C}_{\mathbf{M}}$ *and* $\mathfrak{C}_{\mathbf{H}\to\mathbf{M}}$*. For the former, changing the* `temperature` *amounts to modifying only the corresponding variable* M_1 *and does not affect* M_2*, as evident in Figure 7 (left). Conversely, a change in the* `temperature` *under alignment corresponds also to a change in* `color` *for the variable* M_2*, as depicted in Figure 7 (right). The two interventional effects, hence, do not coincide and* $\mathfrak{C}_{\mathbf{M}}$ *is not an aligned causal abstraction of* $\mathbf{H}$*.*

5. Discussion and Limitations

Our work provides a crisp requirement that machine representations should satisfy to ensure interpretability, namely *alignment* with the human's concept vocabulary. Next, we address important issues arising from this requirement.

5.1. Is Perfect Alignment Sufficient and Necessary?

It is natural to ask whether perfect alignment is a *sufficient* and *necessary* condition for interpretability of machine concepts. Recall that alignment is born out of two desiderata. The first one is that of *subjectivity*: a concept is understandable *to* a particular human observer, with different observers having different expertise and knowledge. This is captured by the human's vocabulary $\mathbf{H}$ in our definition. The second one is that of guaranteeing that *machine and human concepts sharing the same name also share the same semantics*, translated into the desideratum that whenever a human concept changes the human can anticipate how this will change the machine representation. For instance, if the human and the machine see the same picture of a dog, the *human* can easily figure out what concept encodes the notion of "`dog`" and how it would change if they were to delete the dog from the picture. This last point takes into account, at least partially, the limited cognitive processing abilities of human agents.

Is alignment sufficient? Simply ensuring that two agents share aligned representations does not automatically entail that symbolic communication will be successful. For instance, a human observer may misinterpret a machine explanation built out of aligned concepts simply due to inattention, confusion, or information overload. These are all important elements in the equation of interpretability, and we do not intend to dismiss them. The way in which information is *presented* is about as important as the *contents* of the information being conveyed. The problem of designing interfaces that ensure the presentation is engaging and easy to understand is, however, beyond the scope of this paper. This does not impact our core message, that is, that *lack* of alignment can severely hamper communication and that therefore approaches for learning and evaluating conceptual representations should be designed with this requirement in mind.

Is alignment necessary? We also point out that perfect alignment is not strictly *necessary*, for two reasons. First, it is enough that alignment holds only *approximately*. Slight differences in semantics between machine and human concepts are unlikely to have major effects on communication. This is compatible with the empirical observation that people can often successfully communicate even without fully agreeing on the semantics of the words they exchange [63]. In practice, the degree of misalignment, and its impact of the

communication, can be defined and measured, at which point the maximum allowed misalignment becomes an application-specific variable. Second, it may not be necessary that alignment holds *everywhere*. If two agents exchange only a subset of possible messages (e.g., explanations), concepts not appearing in those messages need not be aligned. For instance, ensuring a CBM classifying apples as `ripe` or not to be interpretable only requires the concepts appearing in its explanations to be aligned, and possibly only those values that actually occur in the explanations (e.g., `color` $=$ `red` but not `color` $=$ `blue`). This can be understood as a more lax form of alignment applying only to a certain subset of (values of) the generative factors $\mathbf{g}_\mathcal{I}$, e.g., those related to apples. It is straightforward to relax Definition 6 in this sense by restricting it to a subset of the support of $p^*(\mathbf{G}_\mathcal{I})$ from which the inputs $\mathbf{X}$ are generated, as these constrain the messages that the two agents can exchange.

5.2. Measuring Alignment

While there exist several metrics for measuring interpretability of concepts (discussed in Section 6.4), here we are concerned with techniques for assessing *alignment*.

Considering the relation between alignment and disentanglement (**D1**), one option is to leverage one of the many measures of disentanglement proposed in the literature [64]. The main issues is that most of them provide little information about how simple the map α (**D2**) is and, as such, they cannot be reused as-is. However, for the disentangled case (see Section 4.1), Marconato et al. [19] noted that one can measure alignment using the linear DCI [40]. Essentially, this metric checks whether there exists a *linear regressor* that, given $\mathbf{m}_\mathcal{J}$, can predict $\mathbf{g}_\mathcal{I}$ with high accuracy, such that each M_j is predictive for at most one G_i. In practice, doing so involves collecting a set of annotated pairs $\{(\mathbf{m}_\mathcal{J}, \mathbf{g}_\mathcal{I})\}$, where the m_j's and g_i's are rescaled in $[0, 1]$, and fitting a linear regressor on top of them using L_1 regularization. DCI then considers the (absolute values of the) regressor coefficients $B \in \mathbb{R}^{|\mathcal{J}|\times|\mathcal{I}|}$ and evaluates average dispersion of $B_{j:}$ for each machine representation M_j. In short, if each M_j predicts only a single G_i, and with high accuracy, then linear DCI is maximal. The key insight is that the existence of such a linear map implies both disentanglement (**D1**) *and* monotonicity (**D2**), and therefore also alignment. The main downside is that the converse does not hold, that is, linear DCI cannot account for non-linear monotonic relationships.

The alternative we advocate is that of *decoupling* the measurement of **D1** and **D2**, and to leverage causal notions for the former. **D1** can, for instance, be measured using the *interventional robustness score* (IRS) [41], an empirical version of EMPIDA (Definition 2) that, essentially, measures the average effect of interventions on $\mathbf{G}_\mathcal{I}$ on the machine representation. Alternatives include, for instance, DCI-ES [65], which can better capture the degree by which factors are mixed and the mutual information gap (MIG) [66]. These metrics allow to establish an empirical map π between indices of the human and machine representations, using which it is possible to evaluate **D2** separately. One option is that of evaluating Spearman's rank correlation between the distances:

$$|g_i - g_i'|^2 \quad \text{and} \quad \|\mathbb{E}[\mathbf{M}_\mathcal{J} \mid do(G_i \leftarrow g_i)] - \mathbb{E}[\mathbf{M}_\mathcal{J} \mid do(G_i \leftarrow g_i')]\|_2^2 \tag{19}$$

for interventions g_i and g_i', leaving $\mathbf{G}_{-i}$ fixed, for each $i \in \mathcal{I}$ and multiple traversals (g_i, g_i').

Unfortunately, none of the existing metrics are suited for non-disentangled generative factors $\mathbf{G}_\mathcal{I}$ or human representations $\mathbf{H}$, which are central for alignment in the block-wise (Section 4.4) and general (Section 4.5) cases. Moreover, *alignment* and *block-alignment* share the computational cost and complexity of other disentanglement metrics [41,65,66], since both **D1** and **D2** can be adapted from them. The number of total concept combinations, even in the disentangled case (Definition 1), grows exponentially with the number of concepts k, which requires in practice bounds for the estimation. This is also a noteworthy problem in the disentanglement literature; see, e.g., [64]. We leave an in-depth study of more generally applicable metrics to future work.

5.3. Consequences for Concept-Based Explainers

Recall that CBEs explain the predictions of black-box models by extracting interpretable concepts $\hat{\mathbf{H}}$ from the model's internal representation $\mathbf{M}$ and then evaluating their contribution to the prediction (see Section 3.1). In this case, the requirement is that $\hat{\mathbf{H}}$ is aligned to the human's concept vocabulary $\mathbf{H}$, irrespective of how the former is extracted. Notice that *alignment* is orthogonal to *faithfulness*, in the sense that an aligned representation can be unfaithful to the model, and a faithful representation misaligned with the human. In other words, *alignment is a property of the map from* $\mathbf{H}$ *to* $\hat{\mathbf{H}}$, *while faithfulness is a property of the map between* $\mathbf{M}$ *and* $\hat{\mathbf{H}}$. Evaluating faithfulness can be performed, e.g., via TCAV scores and assessing the degree of linear separation of concepts in the machine representation $\mathbf{M}$. Another approach [36] measures the degree to which $\mathbf{M}$ can be reconducted to a causal decision graph on top of the concepts of interest. Other methods are discussed in Section 6.2.

If the mapping from $\mathbf{M}$ to $\hat{\mathbf{H}}$ is *invertible*, then it is always possible map back and forth, in a lossless manner, from the machine representations $\mathbf{M}$ to the surrogate $\hat{\mathbf{H}}$. This is a solid basis for faithfulness: whatever information is conveyed by an explanation built on $\hat{\mathbf{H}}$ can always be cast in terms of the machine representation itself, and that whatever relation the latter has with the prediction can be mapped in terms of human concepts. The resulting explanation may no longer be simple or understandable, but it still contains all the information of the original message.

In the general case, however, it is non-trivial to find a suitable invertible function. Suppose the user provides the machine with annotated examples $(\mathbf{x}_i, \mathbf{h}_i)$ and that these are used (as is common with supervised CBEs; see Section 6.2) to learn the mapping from $\mathbf{M}$ to $\hat{\mathbf{H}}$. Ensuring that this is invertible requires potentially an enormous amount of examples. To see this, consider a simple case in which the human concepts $\mathbf{H}$ are binary and disentangled and that $\mathbf{M}$ and $\mathbf{H}$ are related by a (possibly complex) invertible mapping that is not an alignment. Even in this ideal case, it might take up to 2^ℓ examples, where ℓ is the dimension of $\mathbf{H}$, to align the two representations, as this involves finding the correct permutation from $\mathbf{M}$ to $\mathbf{H}$. Alignment can help in this regard. In fact, if $\mathbf{M}$ *is* aligned to $\mathbf{H}$, the number of required examples scales as $\mathcal{O}(\ell)$, because a single intervention to each user concept H_i is sufficient to find the corresponding aligned element M_j.

In summary, not only do unaligned (black-box) models imply CBEs require more supervision on the user concepts to acquire a invertible transformation ensuring faithfulness, but it is also likely that the representation $\mathbf{M}$ mixes together the interpretable factors $\mathbf{G}_{\mathcal{I}}$ with the non-interpretable ones $\mathbf{G}_{-\mathcal{I}}$, making it more difficult to extract a concepts $\hat{\mathbf{H}}$ aligned to $\mathbf{H}$.

5.4. Consequences for Concept-Based Models

As discussed in Section 6.1, most CBMs acquire concepts using a variety of heuristics that do not guarantee alignment. To the best of our knowledge, GlanceNets [19] are the only CBM that *explicitly* optimizes for alignment, and as such avoids concept leakage. They do so by combining a variational auto-encoder mapping from the input $\mathbf{X}$ to a machine representation $\mathbf{M} = (\mathbf{M}_{\mathcal{J}}, \mathbf{M}_{-\mathcal{J}})$ where only the first partition is used for prediction. These are computed using a simple linear layer, as is customary. The variational-auto encoder is trained with some amount of concept-level annotations. This encourages both disentanglement [67] *and* monotonicity, and hence alignment, for *in-distribution* data. In turn, this also prevents concept leakage. In order to avoid leakage for *out-of-distribution* data, GlanceNets also implement an *open-set recognition* step [57]. This is responsible for detecting inputs encoding concepts that have never been observed during training. Whenever these are detected, GlanceNets refuse to output a prediction for them, thus avoiding leakage altogether.

From our perspective, GlanceNets have two major downsides. First, they are designed to seek alignment with respect to the generative factors underlying the observations. As we argued, however, interpretability requires alignment with respect to the human's concept vocabulary. Second, GlanceNets require a moderate but non-trivial number of annotations.

How to acquire them from the human observer remains an open problem, as discussed in Section 5.5.

Summarizing, GlanceNets could be repurposed for solving alignment in the disentangled case discussed in Section 4.1 by combining them with a suitable annotation elicitation procedure. They are, however, insufficient to solve disentanglement when the ground-truth concepts are not disentangled, and new solutions will be necessary to tackle these more complex and realistic settings.

5.5. Collecting Human Annotations

Both metrics and learning strategies for alignment require some amount of annotations for the human factors **H**. This is a core requirement related to the subjective nature of interpretability. One option is that of distributing the annotation effort among crowdworkers which, however, is impractical for prediction tasks that require specific types of expertise, like medical diagnosis. An alternative is that of gathering together annotations from different online resources of large language models [68]. Doing so, however, can lead to a lack of completeness (a necessary concept might be missing) and ambiguity (concepts annotations might mix together different views or meanings). This kind of supervision cannot guarantee alignment to a specific human observer.

Reducing the annotation effort for personalized supervision is challenging. Under the assumption that of leveraging generic concept annotations obtained using the above methods to pre-train the concept extractor, and then fine-tune the resulting model using a small amount of personalized annotations. This strategy can save annotation effort as long as the generic annotations contain most of the information necessary to retrieve the observer's concepts. An alternative is to leverage concept-level interactive learning [69,70], to request annotations only for those concepts that are less *aligned*. This is of particular interest for interventions at the concept in CBMs [71,72]. It was also shown that interactive strategies can increase the amount of *disentanglement* during the learning phase [73]. However, how to collect interventional concepts remains an open challenge. Naturally, one might also consider combining these two strategies, that is, interleaving fine-tuning with interactive learning, for additional gains. How to estimate alignment (or some lower bound thereof) in the absence of full concept annotations is, however, an open research question and left to future work.

6. Related Work

While concepts lie at the heart of AI [74], the problem of acquiring *interpretabile* concepts has historically been neglected in representation learning [32]. Recently, concepts have regained popularity in many areas of research, including explainable AI [1], neuro-symbolic AI [75,76], and causality [31,38], yet most concept acquisition strategies developed in these areas are only concerned with task accuracy, rather than interpretability. The presence of users in the framework establishes a connection between our work and cognitive sciences [77,78]. For our purposes, we model the user concepts with variables **H** belonging to their internal structure, whereas other issues arising from the nature of explanations and the subjectivity content are not taken into consideration. This does not exclude that further studies in cognitive science will strengthen the notion of alignment pertaining to individuals' limitations (see, e.g., [79]). Our work builds on causal representation learning [31] that offers a solid basis for capturing some aspects of the mutual understanding between the human and machine, rooted in the definition of alignment.

Next, we briefly overview strategies for acquiring interpretable representations and highlight their shortcomings for properly solving human-interpretable representation learning.

6.1. Unsupervised Approaches

A first group of strategies learn concepts directly from unlabeled data. Well-known theoretical results in deep latent variable models cast doubts on the possibility of acquiring

representations satisfying *any* property of interest, including *disentanglement* and *interpretability*, in a fully unsupervised manner in absence of a strong architectural bias [67,80]. This stems from the fact that, as long as the concept extraction layers are "flexible enough" (i.e., have no strong architectural bias), predictors relying interpretable and uninterpretable concepts can achieve the very same accuracy (or likelihood) on both the training and test sets. As a consequence, *unsupervised strategies that only maximize for accuracy cannot guarantee interpretability unless they are guided by an appropriate bias*. The main challenge is determining what this bias should be.

Several, mutually incompatible alternatives have been proposed. Unsupervised CBEs discover concepts in the space of neuron activations of a target model. One common bias is that concepts can be retrieved by performing a *linear decomposition* of the machine's representation [48]. Specific techniques include k-means [13], principal component analysis [81], and non-negative matrix factorization [14,15]. Concept responsibility is then established via feature attribution methods.

Two common biases used in CBMs are *sparsity* and *orthonormality*. Self-Explainable Neural Networks [16] encourage the former by pairing an autoencoder architecture for extracting concepts from the input together with a (simulatable [21]) task-specific prediction head, and then combining a cross-entropy loss with a penalty term encouraging concepts to have sparse activation patterns. Concept Whitening [27] implements a special bottleneck layer that ensures learned concepts are *orthogonal*, so as to minimize mutual information between them and facilitate acquiring concepts with disjoint semantics, as well as *normalized* within comparable activation ranges. The relationship between sparsity, orthonormality, and interpretability is, however, unclear.

Based on the observation that humans tend to reason in terms of concrete past cases [4], other CBMs constrain concepts to capture salient training examples or parts thereof, i.e., *prototypes*. Methods in this group include Prototype Classification Networks [82], Part-Prototype Networks [17], and many others [83–86]. At a high level, they all memorize one or more prototypes (i.e., points in latent space) that match training examples of their associated class only. Predictions are based on the presence or absence of a match with the learned prototypes. The interpretability of this setup has however been called into question [25,26]. The key issue is the matching step, which is carried out in latent space. The latter is generally underconstrained, meaning that prototypes can end up matching parts of training examples that carry no useful semantics (e.g., arbitrary combinations of foreground and background), as long as doing so yields high training accuracy.

None of these approaches takes the human's own concept vocabulary $\mathbf{H}$ into account.

6.2. Supervised Strategies

A second family of approaches leverages concept *annotations* (or some form of weak supervision). Among supervised CBEs, Net2vec [12] defines linear combinations of convolutional filters, and fits a linear model to decide whether their denoised saliency maps encode a given concept or not, yielding a binary segmentation mask. TCAV [11] defines concepts as directions, or concept-activation vectors (CAVs), in latent space. These are obtained by adapting the parameters of per-concept linear classifiers trained on a separate densely annotated data set to the machine's embedding space. Concept attributions are proportional to the degree by which changing their activations affects the prediction. Zhou et al. [87] also relies on CAVs, but computes explanation by solving an optimization problem. A second group of supervised CBEs makes use of non-linear maps instead [88–90]. For instance, CME [88] uses all activations of the model to learn categorical concepts via semi-supervised multi-task learning, while INN [90] fits a normalizing flow from the machine representation to the concepts so as to guarantee their relationship is bijective. In CBEs, it is also important to estimate the faithfulness of the concepts. TCAV measures the degree of linear separation among concepts [11], Yeh et al. [91] introduced the *completeness score* to evaluate the amount of information the concepts contain about the label, and Fel et al. [48] considered also the CBE *stability*, *fidelity*, and *out-of-distribution discrepancy*.

Supervised CBMs like Concept-Bottleneck Models [18], Concept Whitening [27], and GlanceNets [19], among others [47,92,93] define a loss training penalty, for instance a cross-entropy loss, encouraging the extracted concepts to predict the annotations. Recently, these methods have been extended also to graph neural networks [94,95]. This solution seems straightforward: there is no more direct way than concept supervision to guide the model toward acquiring representations with the intended semantics. It also circumvents the negative theoretical results outlined in Section 6.1.

However, models that accurately match the supervision do not necessarily satisfy content-style separation or allow to have disentangled representations, which, as discussed in Section 4.3, would lead to a non-negligible amount of *concept leakage* [28,29]. In contrast, alignment explicitly takes both properties into account. Another major issue is the supervision itself, which is frequently obtained from general sources rather than from the human observer themselves, meaning the learned concepts may not be aligned to the concept vocabulary of the latter. Two notable exceptions are the interactive concept learning approaches of Lage and Doshi-Velez [69] and of Erculiani et al. [96], which are, however, unconcerned with concept leakage.

To the best of our knowledge, GlanceNets [19] are the only CBM that explicitly optimizes for alignment, and as such, avoid leakage, yet they do so with respect to generative factors rather than human concepts. As discussed in Section 5.4, however, GlanceNets can in principle be adapted to solve human-interpretable representation learning by combining them with a suitable annotation acquisition strategy. We plan to pursue this possibility in future work.

6.3. Disentanglement

Another relevant area of research is that on learning disentangled representations. Here, the goal is to uncover "meaningful", independent factors of variation underlying the data [33,67,97], with the hope that these are also interpretable [32]. Most current learning strategies rely on extensions of variational auto-encoders (VAEs) [66,97–101]. As anticipated in Section 6.1, unless suitable architectural bias is provided, unsupervised learning cannot guarantee the learned representations are disentangled. Motivated by this, follow-up works seek disentanglement via either concept supervision [102], weak supervision [103,104], and other techniques [73,105,106]. Disentanglement, however, is unconcerned with the human's concept vocabulary, and furthermore it is weaker than alignment, in that is does not readily support name transfer.

Independent component analysis (ICA) also seeks to acquire independent factors of variation [107–109]. These assume the generative factors are independent from each other and determine an observation via an injective or invertible map. The objective of ICA is to recover the generative factors from the observations. While the linear case is well understood [107], the non-linear case is arguably more difficult. It was shown that *identifying* the ground truth factor is impossible in the unsupervised setting [110]. This is analogous to the results mentioned in Section 6.1 and, in fact, a formal link between deep latent variable models and identifiability has recently been established [80]. On the positive side, it is possible to show that providing auxiliary supervision on the factors guarantees identification up to permutation and negation, a property known as *strong identifiability*. *Weak identifiability* [111] relaxes it, whereby the generative factors are recovered up to a transformation of the form $A\mathbf{g} + \mathbf{b}$, where $rank(A) \geq \min(\dim \mathbf{G}, \dim \mathbf{M})$ and $\mathbf{M}$ is the machine representation and $\mathbf{b}$ is an offset. Hyvarinen and Morioka [58] also contemplate *identifiability up to element-wise non-linearities*, that is, given by the class of transformations $A\sigma[\mathbf{g}] + \mathbf{b}$, where σ can be a non-linear. If σ is restricted to be monotonic and A is an element-wise transformation, according to condition **D1** in Definition 6, then this form of identifiability matches that of alignment in the disentangled case. However, this formulation refers to identification of the generative factors, while alignment is defined specifically in terms of human concepts. Moreover, we do not assume to the map from human to machine concepts to be injective, nor to be exact.

6.4. Metrics of Concept Quality

Several metrics have been proposed for assessing the quality of extracted concepts and of explanations built on them. Standard measures include accuracy and surrogates thereof [11]; Jaccard similarity [12]; sparsity, stability and ability to reconstruct the model's internal representation [48]; and the degree by which concepts constitute a sufficient statistics for the prediction [91]. We refer to [22] for an overview. These metrics, however, either entirely neglect the role of the human observer, in that concept annotations are either not used or not obtained from the observer themselves, or fail to account for disentanglement and concept leakage. Alignment fills these gaps. Recently, two new metrics have been proposed to measure the concept impurity across individual learned concepts and among sets of representations [112], but the relation with alignment has not been uncovered yet.

There also exist a number of metrics for measuring disentanglement, such as β-VAE score [97], Factor-VAE score [99], mutual information gap [66], DCI [40], and IRS [41]. DCI provides also information about the informativeness of its estimate, and, following [19], it can be repurposed to measure a form of alignment where the μ transformations are linear Definition 6. Suter et al. [41] propose EMPIDA to analyze disentanglement from a causal perspective, upon which we base the construction of alignment. As mentioned in Section 5.2, these metrics can be used to evaluate **D1** in the definition of alignment, and therefore alignment itself when paired with a metric for measuring the complexity of α (**D2**). Their properties are extensively discussed in [64].

6.5. Neuro-Symbolic Architectures

The decomposition between low-level perception—that is, mapping inputs to concepts, also known as *neural predicates* in this setting—and high-level inference outlined in Section 3.1 applies also to many neuro-symbolic (NeSy) models. Examples include DeepProblog [113], Logic Tensor Networks [114], and related architectures [115–125]. The biggest differences between CBMs and NeSy architectures is how they implement the top layer: the former rely on simulatable layers, while the latter on reasoning layers that take prior symbolic knowledge into account and are not necessarily simulatable.

Recent works [126,127] showed that learning a NeSy model consistent with prior knowledge using only label supervision is insufficient to guarantee the neural predicates capture the intended semantics. For instance, it is not uncommon that NeSy architectures attain high prediction accuracy by acquiring neural predicates that encode information about distinct and unrelated concepts. Interpretability of the *neural predicates*, however, also requires alignment, meaning that our results apply to these NeSy architectures as well.

7. Conclusions

Motivated by the growing importance of interpretable representations for both post hoc and ante-hoc explainability, we have introduced and studied the problem of *human-interpretable representation learning*. Our key intuition is that concepts are interpretable only as long as they support symbolic communication with an interested human observer. Based on this, we developed a formal notion of alignment between distributions, rooted in causality, that ensures concepts can support symbolic communication and that applies to both post hoc concept-based explainers and concept-based models. In addition, we clarified the relationship between alignment and the well-known notions of disentanglement, illustrating why the latter is not enough for interpretability, and uncovered a previously unknown link between alignment and concept leakage. Finally, looking at alignment in the most general case, we also unearthed its link to causal abstractions, which further cements the link between interpretability and causality and that we plan to expand on in future work. With this paper, our aim is that of bridging the gap between the human and the algorithmic sides of interpretability, with the hope of providing a solid, mathematical ground on which new research on human-interpretable representation learning can build.

Author Contributions: Conceptualization, all authors; methodology, E.M. and S.T.; supervision, S.T. and A.P.; writing, all authors; funding acquisition, A.P. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge the support of the MUR PNRR project FAIR—Future AI Research (PE00000013) funded by the NextGenerationEU. The research of ST and AP was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs

Appendix A.1. Proof of Proposition 1

The proof requires averaging over the confounds $\mathbf{C}$, encompassing the general case where different Gs may be correlated. To this end, we define the distributions $p(\mathbf{g}) = \mathbb{E}_{\mathbf{C}}[p(\mathbf{G})]$ and $p(\mathbf{G} \mid do(G_i \leftarrow g_i)) = \mathbb{1}\{G_i = g_i\}\mathbb{E}_{\mathbf{C}}[p(\mathbf{G}_{-i} \mid \mathbf{C})]$.

The proof is split into two parts: (i) proving that **D1** implies disentanglement, and (ii) the other way around.

(i) Assume that **D1** holds. Then, the conditional distribution of $\mathbf{M}$ can be written as:

$$p_\theta(\mathbf{m}_\mathcal{J} \mid \mathbf{g}) = \prod_{j\in\mathcal{J}} p_\theta(m_j \mid g_{\pi(j)}) \tag{A1}$$

We proceed to show that Equation (A1) is disentangled in $(\mathbf{G}_\mathcal{I}, \mathbf{M}_\mathcal{J})$. For each $j \in \mathcal{J}$, it holds that the minimum value of $\mathsf{EMPIDA}(G_i, M_j)$ is obtained when $i = \pi(j)$. That is because:

$$\begin{aligned} p_\theta(M_j \mid do(G_{\pi(j)} \leftarrow g_{\pi(j)})) &= \mathbb{E}_{\mathbf{g}_{-\pi(j)}}[p_\theta(M_j \mid g_{\pi(j)})] \\ p_\theta(M_j \mid do(G_{\pi(j)} \leftarrow g_{\pi(j)}, \mathbf{G}_{-\pi(j)} \leftarrow \mathbf{g}_{-\pi(j)})) &= p_\theta(M_j \mid g_{\pi(j)}) \end{aligned} \tag{A2}$$

Note that the first distribution is independent of $\mathbf{g}_{-\pi(j)}$, so it is equivalent to the latter. Hence, $\mathsf{EMPIDA}(G_{\pi(j)}, M_j)$ vanishes $\forall j \in \mathcal{J}$, yielding the claim.

(ii) Let $\mathbf{M}_\mathcal{J}$ now be disentangled with respect to $\mathbf{G}_\mathcal{I}$, that is:

$$\max_{j\in\mathcal{J}} \min_{i\in\mathcal{I}} \mathsf{EMPIDA}(G_i, M_j) = 0 \tag{A3}$$

which is verified *if* it holds that $\min_{i\in\mathcal{I}} \mathsf{EMPIDA}(G_i, M_j) = 0$ for all j. We now proceed by contradiction to show that vanishing EMPIDA is only consistent with **D1**. Suppose there exist at least one $j \in \mathcal{J}$ such that:

$$\alpha(\mathbf{m}_\mathcal{J})_j = \mu_j(\mathbf{g}_{-\mathcal{I}}, N_j) \tag{A4}$$

where $\mathcal{K} \subseteq \mathcal{I}$ containing at least two elements. Therefore, the probability distribution for M_j can be written in general as $p(m_j \mid \mathbf{g}_\mathcal{K})$.

Plugging this condition in the evaluation of EMPIDA, for every $k \in \mathcal{K}$, we obtain:

$$\begin{aligned} p(M_j \mid do(G_k \leftarrow g_k)) &= \mathbb{E}_{\mathbf{g}_{\mathcal{K}\setminus\{k\}}}[p_\theta(m_j \mid g_k, \mathbf{g}_{\mathcal{K}\setminus\{k\}})] \\ p(M_j \mid do(G_k \leftarrow g_k, \mathbf{G}_{-k} \leftarrow \mathbf{g}'_{-k})) &= p_\theta(m_j \mid g_k, \mathbf{g}'_{\mathcal{K}\setminus\{k\}}) \end{aligned} \tag{A5}$$

Then, the two distributions coincide, and EMPIDA is zero, *if* there exists a $k \in \mathcal{K}$ such that all possible interventions $\mathbf{G}_{\mathcal{K}\setminus\{k\}} \leftarrow \mathbf{g}'_{\mathcal{K}\setminus\{k\}}$ do not deviate from the expected distribution, formally:

$$\forall \mathbf{g}'_{\mathcal{K}\setminus\{k\}} \quad p(m_j \mid g_k, \mathbf{g}'_{\mathcal{K}\setminus\{k\}}) = \mathbb{E}_{\mathbf{g}_{\mathcal{K}\setminus\{k\}}} p_\theta(m_j \mid \mathbf{g}_\mathcal{K}) \tag{A6}$$

which holds *if* $p_\theta(m_j \mid \mathbf{g}_k, \mathbf{g}_{\mathcal{K}\setminus\{k\}}) = p_\theta(m_j \mid g_k)$, which is a contradiction. This proves the claim.

Appendix A.2. Proof of Proposition 2

In the following, we adopt the shorthand $\mathbf{m} = \mathbf{m}_\mathcal{J}$, and reintroduce the dependency on $\mathbf{m}_{-\mathcal{J}}$ at the end. First, we show that the maximum of the second term in Λ in Equation (12) coincides with the Shannon entropy of Y:

$$\begin{aligned}\mathcal{L}_r(\gamma) &= \mathbb{E}_{p(\mathbf{x},y)}\big[\log r_\gamma(y)\big] \\ &= \int p(\mathbf{x},y)\log r_\gamma(y)\,\mathrm{d}\mathbf{x}\mathrm{d}y \\ &= \int p(y)\log\frac{r_\gamma(y)p(y)}{p(y)}\,\mathrm{d}y \\ &= -H(Y) - \mathsf{KL}(p(Y)\;||\;r_\gamma(Y))\end{aligned} \tag{A7}$$

where $p(Y)$ denotes the marginal distribution of Y, $H(Y)$ is the Shannon entropy given by $p(Y)$, and KL is the Kullback–Leibler divergence. Since the KL is always non-negative, the previous equation yields the upper bound:

$$\max_\gamma[\mathcal{L}_r(\gamma)] = -H(Y) \tag{A8}$$

We proceed similarly to obtain a lower-bound:

$$\begin{aligned}\mathcal{L}_{CL}(\lambda) &= \int p(\mathbf{x},y)\log\Big(\int q_\lambda(y\mid\mathbf{m})p_\theta(\mathbf{m}\mid\mathbf{x})\,\mathrm{d}\mathbf{m}\Big)\mathrm{d}\mathbf{x}\mathrm{d}y \\ &\geq \int p(\mathbf{x})p_\theta(\mathbf{m}\mid\mathbf{x})p(y\mid\mathbf{x})\log q_\lambda(y\mid\mathbf{m})\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{m}\,\mathrm{d}y \\ &= \int p_\theta(\mathbf{m},y)\log q_\lambda(y\mid\mathbf{m})\,\mathrm{d}\mathbf{m}\mathrm{d}y \\ &= \int p_\theta(\mathbf{m},y)\log\frac{q_\lambda(y\mid\mathbf{m})p_\theta(\mathbf{m})p(y)p_\theta(\mathbf{m},y)}{p_\theta(\mathbf{m})p(y)p_\theta(\mathbf{m},y)}\,\mathrm{d}\mathbf{m}\mathrm{d}y \\ &= \int p_\theta(\mathbf{m},y)\log p(y)\,\mathrm{d}\mathbf{m}\mathrm{d}y + \int p_\theta(\mathbf{m},y)\Big[\log\frac{q_{\lambda,\theta}(\mathbf{m},y)}{p_\theta(\mathbf{m},y)} + \log\frac{p_\theta(\mathbf{m},y)}{p_\theta(\mathbf{m})p(y)}\Big]\,\mathrm{d}\mathbf{m}\mathrm{d}y \\ &= -H(Y) - \mathsf{KL}(p_\theta(\mathbf{M},Y)\;||\;q_{\lambda,\theta}(\mathbf{M},Y)) + \mathrm{I}(\mathbf{M},Y)\end{aligned} \tag{A9}$$

where $p_\theta(\mathbf{m},y) = \int p(\mathbf{x})p_\theta(\mathbf{m}\mid\mathbf{x})p(y\mid\mathbf{x})\,\mathrm{d}\mathbf{x}$, $p_\theta(\mathbf{m})$ is the posterior of the encoding distribution, $q_{\lambda,\theta}(\mathbf{m},y) := q_\lambda(y\mid\mathbf{m})p_\theta(\mathbf{m})$ denotes the joint probability, and $\mathrm{I}(\mathbf{M},Y)$ is the mutual information for the random variables $\mathbf{M}$ and Y, distributed according to $p_\theta(\mathbf{M},y)$. Maximizing the lower-bound implies learning a predictor $q_\lambda(y\mid\mathbf{m})$ that minimizes the KL term. By the previous equation this happens *if* $q_\lambda(y,\mathbf{m})$ matches $p_\theta(\mathbf{m},y)$. Hence, the lower-bound for the first term of Λ becomes:

$$\max_\lambda[\mathcal{L}_{CL}(\lambda)] \geq -H(Y) + \mathrm{I}(\mathbf{M},Y) \tag{A10}$$

Adding this term to the second one shows retrieves the definition of concept leakage and shows that it is lower-bounded by:

$$\Lambda \geq \mathrm{I}(\mathbf{M}_\mathcal{J},Y) \tag{A11}$$

We now proceed deriving the upper-bound for the first term:

$$\begin{aligned}
\mathcal{L}_{CL}(\lambda) &= \int p(\mathbf{x}, y) \log \Big(\int q_\lambda(y \mid \mathbf{m}) p_\theta(\mathbf{m} \mid \mathbf{x}) \, \mathrm{d}\mathbf{m} \Big) \mathrm{d}\mathbf{x} \mathrm{d}y \\
&= \int p(\mathbf{g}_I) q(\mathbf{g}_{-\mathcal{I}}) \Big[\int p(\mathbf{x} \mid \mathbf{g}) p(y \mid \mathbf{g}_{-\mathcal{I}}) \log \Big(\int q_\lambda(y \mid \mathbf{m}) p_\theta(\mathbf{m} \mid \mathbf{x}) \, \mathrm{d}\mathbf{m} \Big) \mathrm{d}\mathbf{x} \, \mathrm{d}y \Big] \, \mathrm{d}\mathbf{g}_{\mathcal{I}} \, \mathrm{d}\mathbf{g}_{-\mathcal{I}} \\
&\leq \int q(\mathbf{g}_{-\mathcal{I}}) \Big[\int p(y \mid \mathbf{g}_{-\mathcal{I}}) \log q_{\lambda,\theta}(y \mid \mathbf{g}_{-\mathcal{I}}) \mathrm{d}y \Big] \, d\mathbf{g}_{-\mathcal{I}} \\
&= \int q(\mathbf{g}_{-\mathcal{I}}) \Big[\int p(y \mid \mathbf{g}_{-\mathcal{I}}) \log \frac{q_{\lambda,\theta}(y \mid \mathbf{g}_{-\mathcal{I}}) p(y) p(y \mid \mathbf{g}_{-\mathcal{I}}) q(\mathbf{g}_{-\mathcal{I}})}{p(y) p(y \mid \mathbf{g}_{-\mathcal{I}}) q(\mathbf{g}_{-\mathcal{I}})} \mathrm{d}y \Big] \, d\mathbf{g}_{-\mathcal{I}} \\
&= \int p(y) \log p(y) \, \mathrm{d}y + \int q(\mathbf{g}_{-\mathcal{I}}) p(y \mid \mathbf{g}_{-\mathcal{I}}) \log \Big[\frac{q_{\lambda,\theta}(y \mid \mathbf{g}_{-\mathcal{I}})}{p(y \mid \mathbf{g}_{-\mathcal{I}})} + \frac{p(y, \mathbf{g}_{-\mathcal{I}})}{p(y) q(\mathbf{g}_{-\mathcal{I}})} \Big] \, \mathrm{d}y \mathrm{d}\mathbf{g}_{-\mathcal{I}} \\
&= -H(Y) - \mathbb{E}_{\mathbf{g}_{-\mathcal{I}} \sim q(\mathbf{g}_{-\mathcal{I}})} [\mathsf{KL}(p(Y \mid \mathbf{G}_{-\mathcal{I}}) \mid\mid q_{\lambda,\theta}(Y \mid \mathbf{G}_{-\mathcal{I}}))] + I(\mathbf{G}_{-\mathcal{I}}, Y)
\end{aligned} \tag{A12}$$

where in the second line, we decomposed $p(\mathbf{x}, y)$ with the data generation process, and in the third line, we made use of Jensen inequality when bringing $\int p(\mathbf{x} \mid \mathbf{g}_{\mathcal{I}}) p(\mathbf{g}_I) \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{g}_{\mathcal{I}}$ in the logarithm, and we denoted with $q_{\lambda,\theta}(y \mid \mathbf{g}_{-\mathcal{I}})$ the conditional distribution obtained by marginalizing over all expectations in the logarithm. Overall, the only part depending on λ appears in the KL term and $I(\mathbf{G}_{-\mathcal{I}}, Y)$ is the mutual information for the probability distribution $p(y \mid \mathbf{g}_{-\mathcal{I}}) q(\mathbf{g}_{-\mathcal{I}})$. Notice that the maximum of the upper-bound for $\mathcal{L}_{CL}(\lambda)$ corresponds to a vanishing KL term and hence the upper-bound for Λ results in:

$$\Lambda \leq I(\mathbf{G}_{-\mathcal{I}}, Y) \tag{A13}$$

Finally, we arrive at the claim:

$$I(\mathbf{M}_{\mathcal{J}}, Y) \leq \Lambda \leq I(\mathbf{G}_{-\mathcal{I}}, Y) \tag{A14}$$

which concludes the proof.

Appendix A.3. Proof of Proposition 3

D1 in Definition 6 entails that the conditional probability of $\mathbf{M}_{\mathcal{J}}$ can be written in general as:

$$p_\theta(\mathbf{m}_{\mathcal{J}} \mid \mathbf{g}) = p_\theta(\mathbf{m}_{\mathcal{J}} \mid \mathbf{g}_{\mathcal{I}}) \tag{A15}$$

The same holds for **D1** in Definition 8. We make use of this fact for deriving a different upper-bound for Λ. We focus only on the first term of Equation (11); the analysis of the second one does not change.

$$\begin{aligned}
\mathcal{L}_{CL}(\lambda) &= \int p(\mathbf{x}, y) \log \Big(\int q_\lambda(y \mid \mathbf{m}_{\mathcal{J}}) p_\theta(\mathbf{m}_{\mathcal{J}} \mid \mathbf{x}) \, \mathrm{d}\mathbf{m}_{\mathcal{J}} \Big) \, \mathrm{d}\mathbf{x} \mathrm{d} \, y \\
&= \int p'(\mathbf{g}) \Big[\int p(\mathbf{x} \mid \mathbf{g}) p(y \mid \mathbf{g}_{-\mathcal{I}}) \log \Big(\int q_\lambda(y \mid \mathbf{m}_{\mathcal{J}}) p_\theta(\mathbf{m}_{\mathcal{J}} \mid \mathbf{x}) \, \mathrm{d}\mathbf{m}_{\mathcal{J}} \Big) \mathrm{d}\mathbf{x} \mathrm{d}y \Big] \, \mathrm{d}\mathbf{g} \\
&\leq \int q(\mathbf{g}_{-\mathcal{I}}) \Big[\int p(y \mid \mathbf{g}_{-\mathcal{I}}) \log \Big(\int q_\lambda(y \mid \mathbf{m}_{\mathcal{J}}) p_\theta(\mathbf{m}_{\mathcal{J}} \mid \mathbf{g}_{\mathcal{I}}) p(\mathbf{g}_{\mathcal{I}}) \, \mathrm{d}\mathbf{m}_{\mathcal{J}} \mathrm{d}\mathbf{g}_{\mathcal{I}} \Big) \mathrm{d}y \Big] \, \mathrm{d}\mathbf{g}_{-\mathcal{I}} \\
&= \int p(y) \log p_{\lambda,\theta}(y) \, \mathrm{d} \, y \\
&= \int p(y) \log \frac{p_{\lambda,\theta}(y) p(y)}{p(y)} \, \mathrm{d} \, y \\
&= -H(Y) - \mathsf{KL}(p(Y) \mid\mid p_{\lambda,\theta}(Y))
\end{aligned} \tag{A16}$$

In the second line, we decomposed the data generation process, and in the third line, we made use of Jensen's inequality to introduce in the logarithm the term $\int p(\mathbf{g}_{\mathcal{I}}) \, \mathrm{d}\mathbf{g}_{\mathcal{I}} \int p(\mathbf{x} \mid \mathbf{g}) \mathrm{d}\mathbf{x}$. The marginalization of $p_\theta(\mathbf{m}_{\mathcal{J}} \mid \mathbf{x})$ with $p(\mathbf{x} \mid \mathbf{g})$ gives $p_\theta(\mathbf{m}_{\mathcal{J}} \mid \mathbf{g})$, that by **D1** reduces to $p_\theta(\mathbf{m}_{\mathcal{J}} \mid \mathbf{g}_{\mathcal{I}})$, hence the term appearing in the third line. In the fourth line, we denoted with $p_{\lambda,\theta}(y) = \int q_\lambda(y \mid \mathbf{m}_{\mathcal{J}}) p_\theta(\mathbf{m}_{\mathcal{J}} \mid \mathbf{g}_{\mathcal{I}}) p(\mathbf{g}_{\mathcal{I}}) \, \mathrm{d}\mathbf{m}_{\mathcal{J}} \mathrm{d}\mathbf{g}_{\mathcal{I}}$ and reduced the first integral

in $p(y)$. Finally, we obtain the upper bound for the first term of Λ, where the maximum implies having a vanishing KL term. Therefore, we have that:

$$\Lambda \leq 0 \tag{A17}$$

Now, since Λ is lower bounded by the mutual information $I(\mathbf{M}_{\mathcal{J}}, Y)$, it cannot be negative and hence must be zero. This concludes the proof.

Appendix A.4. Proof of Corollay 1

The result of the corollary follows from Proposition 3 by considering only the subset of representations $\mathbf{M}_{\mathcal{J}}$ that are not aligned to G_k. Denote them with $\mathbf{M}_{\mathcal{K}}$, where $\mathcal{K} = \{j : \pi(j) \neq k\}$ and set $\bar{\mathbf{G}}_{-\mathcal{I}} = \mathbf{G}_{-\mathcal{I}} \cup G_k$. Then, we have:

$$p_\theta(\mathbf{m}_{\mathcal{K}} \mid \mathbf{g}) = p_\theta(\mathbf{m}_{\mathcal{K}} \mid \mathbf{g}_{\mathcal{I}\setminus\{g_k\}}) \tag{A18}$$

Similarly to Proposition 3, we then obtain that $\Lambda = 0$, i.e., concept leakage vanishes. This proves the claim.

Appendix A.5. Proof of Proposition 4

For a given block $\mathbf{M}_{\mathcal{K}}$ aligned to $\mathbf{G}_{\Pi(\mathcal{K})}$, recall that by **D1** in Definition 8 it holds that:

$$\mathbf{M}_{\mathcal{K}} = \mu_{\mathcal{K}}(\mathbf{G}_{\Pi(\mathcal{K})}, \mathbf{N}_{\mathcal{K}}) \tag{A19}$$

To prove the first claim, we have to show that after intervening on $\mathbf{G}_{\Pi(\mathcal{K})}$ interventions on distinct $\mathbf{G}_{-\Pi(\mathcal{K})}$ do not affect $\mathbf{M}_{\mathcal{K}}$. Fix $do(\mathbf{G}_{\Pi(\mathcal{K})} \leftarrow \mathbf{g}_{\Pi(\mathcal{K})})$. Upon performing a second intervention on the remaining variables $do(\mathbf{G}_{-\Pi(\mathcal{K})} \leftarrow \mathbf{g}_{-\Pi(\mathcal{K})})$, we obtain:

$$p(\mathbf{G}|do(\mathbf{G}_{\Pi(\mathcal{K})} \leftarrow \mathbf{g}_{\Pi(\mathcal{K})}, \mathbf{G}_{-\Pi(\mathcal{K})} \leftarrow \mathbf{g}_{-\Pi(\mathcal{K})})) = \mathbb{1}\Big\{(\mathbf{G}_{\Pi(\mathcal{K})}, \mathbf{G}_{-\Pi(\mathcal{K})}) = (\mathbf{g}_{\Pi(\mathcal{K})}, \mathbf{g}_{-\Pi(\mathcal{K})})\Big\} \tag{A20}$$

By **D1** of Definition 8, it holds that the corresponding probability distribution on $\mathbf{M}_{\mathcal{K}}$ can be written as:

$$p(\mathbf{M}_{\mathcal{K}} \mid do(\mathbf{G}_{\Pi(\mathcal{K})} \leftarrow \mathbf{g}_{\Pi(\mathcal{K})})) = p(\mathbf{M}_{\mathcal{K}} \mid \mathbf{g}_{\Pi(\mathcal{K})}) \tag{A21}$$

which, by a similar argument to Proposition 1, leads to a vanishing $\mathsf{PIDA}(\mathbf{G}_{\Pi(\mathcal{K})}, \mathbf{M}_{\mathcal{K}} \mid \mathbf{g}_{\Pi(\mathcal{K})}, \mathbf{g}_{-\Pi(\mathcal{K})})$ for all possible interventions $do(\mathbf{G}_{-\Pi(\mathcal{K})} \leftarrow \mathbf{g}_{-\Pi(\mathcal{K})})$. This proves that after intervening on $\mathbf{G}_{\Pi(\mathcal{K})}$, arbitrary interventions on $\mathbf{G}_{-\Pi(\mathcal{K})}$ do not affect $\mathbf{M}_{\mathcal{K}}$.

For the second claim, we consider two different intervened values $\mathbf{g}'_{\Pi(\mathcal{K})}$ and $\mathbf{g}''_{\Pi(\mathcal{K})}$ for $\mathbf{G}_{\Pi(\mathcal{K})}$. Recall that by **D2** in Definition 8, it holds that the mean value of $\mathbf{M}_{\mathcal{K}}$ is connected to $\mathbf{G}_{\Pi(\mathcal{K})}$ by an invertible map. Therefore, it holds that:

$$\mathbf{g}'_{\Pi(\mathcal{K})} \neq \mathbf{g}''_{\Pi(\mathcal{K})} \implies \mathbb{E}_{\mathbf{N}_{\mathcal{K}}}[\mu_{\mathcal{K}}(\mathbf{g}'_{\Pi(\mathcal{K})}, \mathbf{N}_{\mathcal{K}})] \neq \mathbb{E}_{\mathbf{N}_{\mathcal{K}}}[\mu_{\mathcal{K}}(\mathbf{g}''_{\Pi(\mathcal{K})}, \mathbf{N}_{\mathcal{K}})] \tag{A22}$$

by invertibility. This concludes the proof.

References

1. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A survey of methods for explaining black box models. *ACM Comput. Surv. (CSUR)* **2018**, *51*, 1–42.
2. Štrumbelj, E.; Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **2014**, *41*, 647–665.
3. Ribeiro, M.T.; Singh, S.; Guestrin, C. "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, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
4. Kim, B.; Khanna, R.; Koyejo, O.O. Examples are not enough, learn to criticize! criticism for interpretability. *Adv. Neural Inf. Process. Syst.* **2016**, *29*.

5. Koh, P.W.; Liang, P. Understanding black-box predictions via influence functions. In Proceedings of the International Conference on Machine Learning, PMLR, Sydney, Australia, 6–11 August 2017; pp. 1885–1894.
6. Ustun, B.; Rudin, C. Supersparse linear integer models for optimized medical scoring systems. *Mach. Learn.* **2016**, *102*, 349–391.
7. Wang, T.; Rudin, C.; Doshi-Velez, F.; Liu, Y.; Klampfl, E.; MacNeille, P. A bayesian framework for learning rule sets for interpretable classification. *J. Mach. Learn. Res.* **2017**, *18*, 2357–2393.
8. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215.
9. Teso, S.; Alkan, Ö.; Stammer, W.; Daly, E. Leveraging Explanations in Interactive Machine Learning: An Overview. *Front. Artif. Intell.* **2023**, *6*, 1066049.
10. Kambhampati, S.; Sreedharan, S.; Verma, M.; Zha, Y.; Guan, L. Symbols as a lingua franca for bridging human-ai chasm for explainable and advisable ai systems. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 28 February–1 March 2022; Volume 36, pp. 12262–12267.
11. Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F. Interpretability beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm, Sweden, 10–15 July 2018; pp. 2668–2677.
12. Fong, R.; Vedaldi, A. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Utah, USA, 18–22 June 2018; pp. 8730–8738.
13. Ghorbani, A.; Abid, A.; Zou, J. Interpretation of neural networks is fragile. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, January 27–February 1, 2019; Volume 33, pp. 3681–3688.
14. Zhang, R.; Madumal, P.; Miller, T.; Ehinger, K.A.; Rubinstein, B.I. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually , 2–9 February 2021; Volume 35, pp. 11682–11690.
15. Fel, T.; Picard, A.; Bethune, L.; Boissin, T.; Vigouroux, D.; Colin, J.; Cadène, R.; Serre, T. Craft: Concept recursive activation factorization for explainability. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14-19 June 2020; pp. 2711–2721.
16. Alvarez-Melis, D.; Jaakkola, T.S. Towards robust interpretability with self-explaining neural networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, Canada, 3-8 December 2018; pp. 7786–7795.
17. Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; Su, J.K. This Looks Like That: Deep Learning for Interpretable Image Recognition. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 8930–8941.
18. Koh, P.W.; Nguyen, T.; Tang, Y.S.; Mussmann, S.; Pierson, E.; Kim, B.; Liang, P. Concept bottleneck models. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13-18 July 2020; pp. 5338–5348.
19. Marconato, E.; Passerini, A.; Teso, S. GlanceNets: Interpretabile, Leak-proof Concept-based Models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 21212–21227.
20. Espinosa Zarlenga, M.; Barbiero, P.; Ciravegna, G.; Marra, G.; Giannini, F.; Diligenti, M.; Shams, Z.; Precioso, F.; Melacci, S.; Weller, A.; et al. Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 21400–21413.
21. Lipton, Z.C. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, *16*, 31–57.
22. Schwalbe, G. Concept embedding analysis: A review. *arXiv* **2022**. arXiv:2203.13909.
23. Stammer, W.; Schramowski, P.; Kersting, K. Right for the Right Concept: Revising Neuro-Symbolic Concepts by Interacting with their Explanations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 3619–3629.
24. Bontempelli, A.; Teso, S.; Giunchiglia, F.; Passerini, A. Concept-level debugging of part-prototype networks. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
25. Hoffmann, A.; Fanconi, C.; Rade, R.; Kohler, J. This Looks Like That... Does it? Shortcomings of Latent Space Prototype Interpretability in Deep Networks. *arXiv* **2021**. arXiv:2105.02968.
26. Xu-Darme, R.; Quénot, G.; Chihani, Z.; Rousset, M.C. Sanity Checks for Patch Visualisation in Prototype-Based Image Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 18–22 June 2023; pp. 3690–3695.
27. Chen, Z.; Bei, Y.; Rudin, C. Concept whitening for interpretable image recognition. *Nat. Mach. Intell.* **2020**, *2*, 772–782.
28. Margeloiu, A.; Ashman, M.; Bhatt, U.; Chen, Y.; Jamnik, M.; Weller, A. Do Concept Bottleneck Models Learn as Intended? *arXiv* **2021**. arXiv:2105.04289.
29. Mahinpei, A.; Clark, J.; Lage, I.; Doshi-Velez, F.; Pan, W. Promises and pitfalls of black-box concept learning models. In Proceedings of the International Conference on Machine Learning: Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI, Virtual, 8–9 February 2021; Volume 1, pp. 1–13.
30. Silver, D.L.; Mitchell, T.M. The Roles of Symbols in Neural-based AI: They are Not What You Think! *arXiv* **2023**. arXiv:2304.13626.
31. Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N.R.; Kalchbrenner, N.; Goyal, A.; Bengio, Y. Toward causal representation learning. *Proc. IEEE* **2021**, *109*, 612–634.

32. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828.
33. Higgins, I.; Amos, D.; Pfau, D.; Racaniere, S.; Matthey, L.; Rezende, D.; Lerchner, A. Towards a definition of disentangled representations. *arXiv* **2018**. arXiv:1812.02230.
34. Beckers, S.; Halpern, J.Y. Abstracting causal models. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 33, pp. 2678–2685.
35. Beckers, S.; Eberhardt, F.; Halpern, J.Y. Approximate causal abstractions. In Proceedings of the Uncertainty in Artificial Intelligence, PMLR, Online, 3–6 August 2020; pp. 606–615.
36. Geiger, A.; Wu, Z.; Potts, C.; Icard, T.; Goodman, N.D. Finding alignments between interpretable causal variables and distributed neural representations. *arXiv* **2023**. arXiv:2303.02536.
37. Lockhart, J.; Marchesotti, N.; Magazzeni, D.; Veloso, M. Towards learning to explain with concept bottleneck models: mitigating information leakage. *arXiv* **2022**. arXiv:2211.03656.
38. Pearl, J. *Causality*; Cambridge University Press: Cambridge, UK, 2009.
39. Peters, J.; Janzing, D.; Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*; MIT Press: Cambridge, MA, USA, 2017.
40. Eastwood, C.; Williams, C.K. A framework for the quantitative evaluation of disentangled representations. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
41. Suter, R.; Miladinovic, D.; Schölkopf, B.; Bauer, S. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 6056–6065.
42. Reddy, A.G.; Balasubramanian, V.N. On causally disentangled representations. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 28 February–1 March 2022; Volume 36, pp. 8089–8097.
43. von Kügelgen, J.; Sharma, Y.; Gresele, L.; Brendel, W.; Schölkopf, B.; Besserve, M.; Locatello, F. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In Proceedings of the 35nd International Conference on Neural Information Processing Systems, Online, 6–14 December 2021.
44. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*; MIT Press: Cambridge, MA, USA, 2009.
45. Yang, Y.; Panagopoulou, A.; Zhou, S.; Jin, D.; Callison-Burch, C.; Yatskar, M. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 18–22 June 2023; pp. 19187–19197.
46. Bontempelli, A.; Giunchiglia, F.; Passerini, A.; Teso, S. Toward a Unified Framework for Debugging Gray-box Models. In Proceedings of the The AAAI-22 Workshop on Interactive Machine Learning, Online, 28 February 2022.
47. Zarlenga, M.E.; Pietro, B.; Gabriele, C.; Giuseppe, M.; Giannini, F.; Diligenti, M.; Zohreh, S.; Frederic, P.; Melacci, S.; Adrian, W.; et al. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Needham, MA, USA, 2022; Volume 35, pp. 21400–21413.
48. Fel, T.; Boutin, V.; Moayeri, M.; Cadène, R.; Bethune, L.; andéol, L.; Chalvidal, M.; Serre, T. A Holistic Approach to Unifying Automatic Concept Extraction and Concept Importance Estimation. *arXiv* 2023. arXiv:2306.07304.
49. Teso, S. Toward Faithful Explanatory Active Learning with Self-explainable Neural Nets. In Proceedings of the Workshop on Interactive Adaptive Learning (IAL 2019), 2019; pp. 4–16. Available online: https://ceur-ws.org/Vol-2444/ialatecml_paper1.pdf (accessed on 9 September 2023).
50. Pfau, J.; Young, A.T.; Wei, J.; Wei, M.L.; Keiser, M.J. Robust semantic interpretability: Revisiting concept activation vectors. *arXiv* **2021**. arXiv:2104.02768.
51. Gabbay, A.; Cohen, N.; Hoshen, Y. An image is worth more than a thousand words: Towards disentanglement in the wild. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9216–9228.
52. Matthey, L.; Higgins, I.; Hassabis, D.; Lerchner, A. dSprites: Disentanglement Testing Sprites Dataset. 2017. Available online: https://github.com/deepmind/dsprites-dataset/ (accessed on 9 September 2023).
53. Havasi, M.; Parbhoo, S.; Doshi-Velez, F. Addressing Leakage in Concept Bottleneck Models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 23386–23397.
54. Cover, T.M. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 1999.
55. Montero, M.L.; Ludwig, C.J.; Costa, R.P.; Malhotra, G.; Bowers, J. The role of disentanglement in generalisation. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 30 April 2020.
56. Montero, M.; Bowers, J.; Ponte Costa, R.; Ludwig, C.; Malhotra, G. Lost in Latent Space: Examining failures of disentangled models at combinatorial generalisation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 10136–10149.
57. Sun, X.; Yang, Z.; Zhang, C.; Ling, K.V.; Peng, G. Conditional gaussian distribution learning for open set recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 14–19 June 2020; pp. 13480–13489.
58. Hyvarinen, A.; Morioka, H. Nonlinear ICA of temporally dependent stationary sources. In Proceedings of the Artificial Intelligence and Statistics, PMLR, Ft. Lauderdale, FL, USA, 20–22 April 2017; pp. 460–469.

59. Khemakhem, I.; Monti, R.P.; Kingma, D.P.; Hyvärinen, A. ICE-BeeM: Identifiable Conditional Energy-Based Deep Models Based on Nonlinear ICA. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Online, 6–12 December 2020.
60. Rubenstein, P.K.; Weichwald, S.; Bongers, S.; Mooij, J.M.; Janzing, D.; Grosse-Wentrup, M.; Schölkopf, B. Causal consistency of structural equation models. *arXiv* **2017**. arXiv:1707.00819.
61. Zennaro, F.M. Abstraction between structural causal models: A review of definitions and properties. *arXiv* **2022**. arXiv:2207.08603.
62. Geiger, A.; Potts, C.; Icard, T. Causal Abstraction for Faithful Model Interpretation. *arXiv* **2023**. arXiv:2301.04709.
63. Marti, L.; Wu, S.; Piantadosi, S.T.; Kidd, C. Latent diversity in human concepts. *Open Mind* **2023**, *7*, 79–92.
64. Zaidi, J.; Boilard, J.; Gagnon, G.; Carbonneau, M.A. Measuring disentanglement: A review of metrics. *arXiv* **2020**. arXiv:2012.09276.
65. Eastwood, C.; Nicolicioiu, A.L.; Von Kügelgen, J.; Kekić, A.; Träuble, F.; Dittadi, A.; Schölkopf, B. DCI-ES: An Extended Disentanglement Framework with Connections to Identifiability. *arXiv* **2022**. arXiv:2210.00364.
66. Chen, R.T.; Li, X.; Grosse, R.; Duvenaud, D. Isolating sources of disentanglement in VAEs. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, Canada, 3–8 December 2018; pp. 2615–2625.
67. Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 4114–4124.
68. Oikarinen, T.; Das, S.; Nguyen, L.M.; Weng, T.W. Label-free Concept Bottleneck Models. In Proceedings of the ICLR, Virtual, 25 April 2022.
69. Lage, I.; Doshi-Velez, F. Learning Interpretable Concept-Based Models with Human Feedback. *arXiv* **2020**. arXiv:2012.02898.
70. Chauhan, K.; Tiwari, R.; Freyberg, J.; Shenoy, P.; Dvijotham, K. Interactive concept bottleneck models. In Proceedings of the AAAI, Washington, DC, USA, 7–14 February 2023.
71. Steinmann, D.; Stammer, W.; Friedrich, F.; Kersting, K. Learning to Intervene on Concept Bottlenecks. *arXiv* **2023**. arXiv:2308.13453.
72. Zarlenga, M.E.; Collins, K.M.; Dvijotham, K.; Weller, A.; Shams, Z.; Jamnik, M. Learning to Receive Help: Intervention-Aware Concept Embedding Models. *arXiv* **2023**. arXiv:2309.16928.
73. Stammer, W.; Memmel, M.; Schramowski, P.; Kersting, K. Interactive disentanglement: Learning concepts by interacting with their prototype representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10317–10328.
74. Muggleton, S.; De Raedt, L. Inductive logic programming: Theory and methods. *J. Log. Program.* **1994**, *19*, 629–679.
75. De Raedt, L.; Dumancic, S.; Manhaeve, R.; Marra, G. From Statistical Relational to Neuro-Symbolic Artificial Intelligence. In Proceedings of the IJCAI, Yokohama, Japan, 11–17 July 2020.
76. Holzinger, A.; Saranti, A.; Angerschmid, A.; Finzel, B.; Schmid, U.; Mueller, H. Toward human-level concept learning: Pattern benchmarking for AI algorithms. *Patterns* **2023**, *4*, 100788.
77. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.* **2019**, *267*, 1–38.
78. Cabitza, F.; Campagner, A.; Malgieri, G.; Natali, C.; Schneeberger, D.; Stoeger, K.; Holzinger, A. Quod erat demonstrandum?-Towards a typology of the concept of explanation for the design of explainable AI. *Expert Syst. Appl.* **2023**, *213*, 118888.
79. Ho, M.K.; Abel, D.; Correa, C.G.; Littman, M.L.; Cohen, J.D.; Griffiths, T.L. People construct simplified mental representations to plan. *Nature* **2022**, *606*, 129–136.
80. Khemakhem, I.; Kingma, D.; Monti, R.; Hyvarinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, Online, 26–28 August 2020; pp. 2207–2217.
81. Graziani, M.; Nguyen, A.P.; O'Mahony, L.; Müller, H.; Andrearczyk, V. Concept discovery and dataset exploration with singular value decomposition. In Proceedings of the ICLR 2023 Workshop on Pitfalls of Limited Data and Computation for Trustworthy ML, Kigali, Rwanda, 5 May 2023.
82. Li, O.; Liu, H.; Chen, C.; Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In Proceedings of the AAAI Conference on Artificial Intelligence, Orleans, LA, USA, 2–7 February 2018.
83. Rymarczyk, D.; Struski, L.; Tabor, J.; Zieliński, B. ProtoPShare: Prototypical Parts Sharing for Similarity Discovery in Interpretable Image Classification. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore, 14–18 August2021; pp. 1420–1430.
84. Nauta, M.; van Bree, R.; Seifert, C. Neural Prototype Trees for Interpretable Fine-grained Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 14933–14943.
85. Singh, G.; Yow, K.C. These do not look like those: An interpretable deep learning model for image recognition. *IEEE Access* **2021**, *9*, 41482–41493.
86. Davoudi, S.O.; Komeili, M. Toward Faithful Case-based Reasoning through Learning Prototypes in a Nearest Neighbor-friendly Space. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
87. Zhou, B.; Sun, Y.; Bau, D.; Torralba, A. Interpretable basis decomposition for visual explanation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 119–134.

88. Kazhdan, D.; Dimanov, B.; Jamnik, M.; Liò, P.; Weller, A. Now you see me (CME): Concept-based model extraction. *arXiv* **2020**. arXiv:2010.13233.
89. Gu, J.; Tresp, V. Semantics for global and local interpretation of deep neural networks. *arXiv* **2019**. arXiv:1910.09085.
90. Esser, P.; Rombach, R.; Ommer, B. A disentangling invertible interpretation network for explaining latent representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 9223–9232.
91. Yeh, C.K.; Kim, B.; Arik, S.; Li, C.L.; Pfister, T.; Ravikumar, P. On completeness-aware concept-based explanations in deep neural networks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 20554–20565.
92. Yuksekgonul, M.; Wang, M.; Zou, J. Post-hoc Concept Bottleneck Models. *arXiv* **2022**. arXiv:2205.15480.
93. Sawada, Y.; Nakamura, K. Concept Bottleneck Model with Additional Unsupervised Concepts. *IEEE Access* **2022**, *10*, 41758–41765.
94. Magister, L.C.; Kazhdan, D.; Singh, V.; Liò, P. Gcexplainer: Human-in-the-loop concept-based explanations for graph neural networks. *arXiv* **2021**. arXiv:2107.11889.
95. Finzel, B.; Saranti, A.; Angerschmid, A.; Tafler, D.; Pfeifer, B.; Holzinger, A. Generating explanations for conceptual validation of graph neural networks: An investigation of symbolic predicates learned on relevance-ranked sub-graphs. *KI-Künstliche Intell.* **2022**, *36*, 271–285.
96. Erculiani, L.; Bontempelli, A.; Passerini, A.; Giunchiglia, F. Egocentric Hierarchical Visual Semantics. *arXiv* **2023**. arXiv:2305.05422.
97. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
98. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. In Proceedings of the International Conference on Machine Learning, PMLR, Beijing, China, 22–24 June 2014.
99. Kim, H.; Mnih, A. Disentangling by factorising. In Proceedings of the International Conference on Machine Learning, PMLR, Stockholm Sweden, 10–15 July 2018; pp. 2649–2658.
100. Esmaeili, B.; Wu, H.; Jain, S.; Bozkurt, A.; Siddharth, N.; Paige, B.; Brooks, D.H.; Dy, J.; Meent, J.W. Structured disentangled representations. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, PMLR, Naha, Okinawa, Japan, 16–18 April 2019; pp. 2525–2534.
101. Rhodes, T.; Lee, D. Local Disentanglement in Variational Auto-Encoders Using Jacobian L_1 Regularization. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 22708–22719.
102. Locatello, F.; Tschannen, M.; Bauer, S.; Rätsch, G.; Schölkopf, B.; Bachem, O. Disentangling Factors of Variations Using Few Labels. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 30 April 2020.
103. Shu, R.; Chen, Y.; Kumar, A.; Ermon, S.; Poole, B. Weakly Supervised Disentanglement with Guarantees. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 30 April 2020.
104. Locatello, F.; Poole, B.; Rätsch, G.; Schölkopf, B.; Bachem, O.; Tschannen, M. Weakly-supervised disentanglement without compromises. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 6348–6359.
105. Lachapelle, S.; Rodriguez, P.; Sharma, Y.; Everett, K.E.; Le Priol, R.; Lacoste, A.; Lacoste-Julien, S. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In Proceedings of the Conference on Causal Learning and Reasoning, PMLR, Eureka, CA, USA, 11–13 April 2022; pp. 428–484.
106. Horan, D.; Richardson, E.; Weiss, Y. When Is Unsupervised Disentanglement Possible? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 5150–5161.
107. Comon, P. Independent component analysis, a new concept? *Signal Process.* **1994**, *36*, 287–314.
108. Hyvärinen, A.; Karhunen, J.; Oja, E. *Independent Component Analysis, Adaptive and Learning Systems for Signal Processing, Communications, and Control*; John Wiley Sons, Inc.: Hoboken, NJ, USA, 2001; Volume 1, pp. 11–14.
109. Naik, G.R.; Kumar, D.K. An overview of independent component analysis and its applications. *Informatica* **2011**, *35*, 63–81.
110. Hyvärinen, A.; Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Netw.* **1999**, *12*, 429–439.
111. Buchholz, S.; Besserve, M.; Schölkopf, B. Function classes for identifiable nonlinear independent component analysis. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 16946–16961.
112. Zarlenga, M.E.; Barbiero, P.; Shams, Z.; Kazhdan, D.; Bhatt, U.; Weller, A.; Jamnik, M. Towards Robust Metrics for Concept Representation Evaluation. *arXiv* **2023**. arXiv:2301.10367.
113. Manhaeve, R.; Dumancic, S.; Kimmig, A.; Demeester, T.; De Raedt, L. DeepProbLog: Neural Probabilistic Logic Programming. *Adv. Neural Inf. Process. Syst.* **2021**, *31*, 3753–3763.
114. Donadello, I.; Serafini, L.; Garcez, A.D. Logic tensor networks for semantic image interpretation. *arXiv* **2017**. arXiv:1705.08968.
115. Diligenti, M.; Gori, M.; Sacca, C. Semantic-based regularization for learning and inference. *Artif. Intell.* **2017**, *244*, 143–165.
116. Fischer, M.; Balunovic, M.; Drachsler-Cohen, D.; Gehr, T.; Zhang, C.; Vechev, M. Dl2: Training and querying neural networks with logic. In Proceedings of the International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019; pp. 1931–1941.
117. Giunchiglia, E.; Lukasiewicz, T. Coherent Hierarchical Multi-label Classification Networks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9662–9673.

118. Yang, Z.; Ishay, A.; Lee, J. NeurASP: Embracing neural networks into answer set programming. In Proceedings of the IJCAI, Long Beach, CA, USA, 9–15 June 2019.
119. Huang, J.; Li, Z.; Chen, B.; Samel, K.; Naik, M.; Song, L.; Si, X. Scallop: From Probabilistic Deductive Databases to Scalable Differentiable Reasoning. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 25134–25145.
120. Marra, G.; Kuželka, O. Neural markov logic networks. In Proceedings of the Uncertainty in Artificial Intelligence, Online, 27–30 July 2021.
121. Ahmed, K.; Teso, S.; Chang, K.W.; Van den Broeck, G.; Vergari, A. Semantic Probabilistic Layers for Neuro-Symbolic Learning. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 29944–29959.
122. Misino, E.; Marra, G.; Sansone, E. VAEL: Bridging Variational Autoencoders and Probabilistic Logic Programming. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 4667–4679.
123. Winters, T.; Marra, G.; Manhaeve, R.; De Raedt, L. DeepStochLog: Neural Stochastic Logic Programming. In Proceedings of the AAAI, virtually, 22 February–1 March 2022.
124. van Krieken, E.; Thanapalasingam, T.; Tomczak, J.M.; van Harmelen, F.; Teije, A.T. A-NeSI: A Scalable Approximate Method for Probabilistic Neurosymbolic Inference. *arXiv* **2022**. arXiv:2212.12393.
125. Ciravegna, G.; Barbiero, P.; Giannini, F.; Gori, M.; Lió, P.; Maggini, M.; Melacci, S. Logic explained networks. *Artif. Intell.* **2023**, *314*, 103822.
126. Marconato, E.; Bontempo, G.; Ficarra, E.; Calderara, S.; Passerini, A.; Teso, S. Neuro-Symbolic Continual Learning: Knowledge, Reasoning Shortcuts and Concept Rehearsal. In Proceedings of the 40th International Conference on Machine Learning (ICML'23), Honolulu, HI , USA, 23–29 July 2023;Volume 202, pp. 23915–23936.
127. Marconato, E.; Teso, S.; Vergari, A.; Passerini, A. Not All Neuro-Symbolic Concepts Are Created Equal: Analysis and Mitigation of Reasoning Shortcuts. In Proceedings of the Thirty-Seventh Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023.