Position: Actionable Interpretability Must Be Defined in Terms of Symmetries

Pietro Barbiero [1] **Mateo Espinosa Zarlenga** [2] **Francesco Giannini** [3] **Alberto Termine** [4] **Filippo Bonchi** [3] **Mateja Jamnik** [5] **Giuseppe Marra** [6]

Abstract

This paper argues that interpretability research in Artificial Intelligence (AI) is fundamentally ill-posed as existing definitions of interpretability are not *actionable*: they fail to provide formal principles from which concrete modelling and inferential rules can be derived. We posit that for a definition of interpretability to be actionable, it must be given in terms of *symmetries*. We hypothesise that four symmetries suffice to (i) motivate core interpretability properties, (ii) characterize the class of interpretable models, and (iii) derive a unified formulation of interpretable inference (e.g., alignment, interventions, and counterfactuals) as a form of Bayesian inversions.

1. Introduction

Recent years have seen a surge in interpretable models whose decisions can be easily understood by humans. These models now offer a performance comparable to that of powerful black-box models like Deep Neural Networks (DNNs) (Alvarez-Melis & Jaakkola, 2018; Chen et al., 2019; Espinosa Zarlenga et al., 2022), and are increasingly employed to diagnose errors, ensure fairness, and comply with legal standards (Lee et al., 2021; Meng et al., 2022).

Thus, while of primary importance, we argue that current research in interpretable Artificial Intelligence (AI) is ill-posed since existing foundational efforts (Sec. 5), e.g., (Kim et al., 2016; Biran & Cotton, 2017; Doshi-Velez & Kim, 2017; Lipton, 2018; Miller, 2019; Watson & Floridi, 2021; Facchini & Termine, 2021; Giannini et al., 2024; Tull et al., 2024), fail to formalise general principles from which concrete modelling design choices can be derived.

In the spirit of the Erlangen Program (Klein, 1893; Bronstein et al., 2021), we posit that for a definition of interpretability to be *actionable* it must be formulated as a set of *symmetries*, that is, interpretability "first principles" must be characterised in terms of transformations under which essential interpretability properties remain invariant. We hypothesise that four such symmetries are sufficient to formalise the first principles of interpretability (Sec. 2), from which all further properties follow:

- **Inference equivariance.** A model is interpretable if a user can correctly predict the model outputs by simulating its decision-making process.
- **Information invariance.** An interpretable model should retain only input information that is sufficient for the task, discarding irrelevant details (e.g., an individual pixel intensity is unnecessary for classifying cats vs. dogs).
- **Concept-closure invariance.** The representations used by an interpretable model should correspond to units of information used by humans, so that model variables have stable meaning under translation into human terms (e.g., "red" in the model matches "red" for a human).
- **Structural invariance.** A model is interpretable if it is drawn from a hypothesis class the user can reason about, that is, if a user can only reason in linear terms, the model must satisfy linearity.

To concretely assess these symmetries, we examine which properties from the literature can be derived from them, which methods satisfy the corresponding conditions, and which models fail to do so and therefore cannot be considered interpretability-complete. We then show how these formal principles induce the construction of the category of interpretable models, that is, the recipe specifying the ingredients and the elementary cooking operations to build any interpretable model (Sec. 3). Finally, we show how the constructed category allows the unification of alignment, intervention, and counterfactual inference as a form of Bayesian inversion (Sec. 4). The main ideas of the paper can be understood in under a minute by reading only the highlighted text boxes.

[1]IBM Research [2]University of Oxford [3]University of Pisa [4]SUPSI [5]University of Cambridge [6]KU Leuven. Correspondence to: Pietro Barbiero <barbiero@tutanota.com>.

2. Interpretability Symmetries

In this section we introduce four symmetries and hypothesise that their satisfaction characterizes the class of models that are interpretable for a given human user. Inspired by mathematical foundations of explainable AI (Giannini et al., 2024; Tull et al., 2024; Colombini et al., 2025; Fioravanti et al., 2025), we rely on category theory[1] as an expressive framework to characterize (i) human users and (ii) the class of models that are interpretable.

Setup. Following Murphy (2023), we adopt a probabilistic perspective. Categorically, this viewpoint is captured by *Markov categories* Stoch (Fritz, 2020), which provide a general and compositional language for probabilistic models. **Human:** We denote by h a human user and by Hm the category of models (that is, the hypothesis space) of the user's "*mental models*" (Johnson-Laird, 1983; Kulesza et al., 2013; Kim et al., 2018). Formally, Hm is characterised by a set of user-dependent structural properties (such as linearity, simplicity, or monotonicity) which are assumed to be given. **Model:** We then define $\mathsf{Im}_{[h]}$ as the category of models that are interpretable for h, and consider both Hm and $\mathsf{Im}_{[h]}$ to be subcategories of a common Markov category Stoch. **Notation:** Given a sample space Ω, we introduce a random variable $X : \Omega \to \mathcal{X}$, where $X(\omega)$ denotes a feature-based representation of the object associated with outcome $\omega \in \Omega$, a random variable Y, and a probabilistic model $P_{Y|X} : \mathcal{X} \to \mathbb{P}(Y)$ that assigns to each value $x \in \mathcal{X}$ a probability distribution on Y, i.e. $P_{Y|X}(x) = P(Y|X = x)$. Probabilistic inference evaluates the conditional model $P_{Y|X}$ on any input $x \in \mathcal{X}$ to obtain a predictive distribution over Y.

2.1. Symmetry I: Inference Equivariance

Most works in the interpretability literature define their subject matter in informal terms. For instance, Kim et al. (2016) and Miller (2019) suggest that *a method is interpretable if a user can correctly and efficiently predict the method's results* and *the cause of such results*. Closely related, Biran & Cotton (2017) and Murdoch et al. (2019) argue that a *system is interpretable if a human is able to internally simulate and reason about its entire decision-making process (i.e. how a trained model produces an output for an arbitrary input)*. Similar views connecting the notion of interpretability to how *intelligible* a model's inference is have been proposed by Lou et al. (2012) and Ribeiro et al. (2016). While influential, these definitions remain largely descriptive. Our first goal will be then to answer the following question: **(RQ1)** *How can we formally bring together the informal descriptions of interpretability?*

[1]While not essential to understand this paper, basic notions in category theory related to this work are presented by Jacobs et al. (2025); Tull et al. (2024).

To answer this question, we introduce an example concretely showing what it would mean for a user to "simulate a model" and "predict the model's results". We consider a finite space of values for X, Y (with $X = (X_1, X_2)$) so that we can give a complete representation of the model's behaviour via a conditional probability table:

Table 1: Tabular representation of a model $P_{Y|X}$.

	X_1	X_2	$P_{Y\|X}(Y = 1 \mid X_1 = x_1, X_2 = x_2)$
	0	1	1
	0	0	1
	1	0	0

We now fix a *translation map* that associates the model's objects with representations meaningful to a given user h, that is, $\tau : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}_{[h]} \times \mathcal{Y}_{[h]}$. For example, τ may map the objects $\mathcal{X}_1, \mathcal{X}_2, \mathcal{Y}$ to the user variables $\mathcal{X}_{\texttt{one}}, \mathcal{X}_{\texttt{red}}, \mathcal{Y}_{\texttt{even}}$. Given the original model and the translation map τ, the user h can *simulate* the model if the user can build a "mental model" $P^{[h]}_{Y|X}$ in Hm whose predictions match the translated outputs of $P_{Y|X}$. In practice, we can check whether human can simulate the model as follows: (1) we can first translate the input features X into a human input space $X_{[h]}$ and predict the target using the human model $P^{[h]}_{Y|X}$, or (2) we can first infer a label using $P_{Y|X}$ and then translate the result. Visually, we represent these two paths as:

$$
\begin{tikzcd}
(X_1, X_2) = (0,0) \arrow[r, "\text{unknown model } P_{Y|X}"] \arrow[d, "\text{``translate''}"'] & Y = 1 \arrow[d, "\text{``translate''}"] \\
(X_{\texttt{one}}, X_{\texttt{red}}) = (\text{no}, \text{no}) \arrow[r, "\text{human model } P^{[h]}_{Y|X}"'] & Y_{\texttt{even}} = \text{yes}
\end{tikzcd}
$$

If this diagram commutes for *any* input $x \in \mathcal{X}$ (i.e., if we reach the same result following different paths), then it means that the model $P_{Y|X}$ can be "simulated" by the human model $P^{[h]}_{Y|X}$. Inspired by Marconato et al. (2023), we refer to this principle as *inference equivariance*:

Symmetry 1. (**Inference Equivariance**) Inference $P_{Y|X}(Y \mid X = x)$ is *equivariant* w.r.t. a reference Hm under a translation $\tau : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}_{[h]} \times \mathcal{Y}_{[h]}$ if and only if it exists a model $P^{[h]}_{Y|X} : \mathcal{X}_{[h]} \to \mathbb{P}(Y)$ in Hm such that the following diagram commutes:

$$
\begin{tikzcd}
\mathcal{X} \arrow[r, "P_{Y|X}"] \arrow[d, "\tau"'] & \mathcal{Y} \arrow[d, "\tau"] \\
\mathcal{X}_{[h]} \arrow[r, "P^{[h]}_{Y|X}"'] & \mathcal{Y}_{[h]}
\end{tikzcd}
$$

Why we need more symmetries While this symmetry subsumes prior definitions of interpretability by Kim et al. (2016); Miller (2019); Biran & Cotton (2017); Murdoch et al.

(2019), it also emphasises that several important questions remain open:

C1 **Naively verifying inference equivariance is intractable.** To guarantee that inference equivariance always holds for any $x \in \mathcal{X}$, we need a table with $\mathcal{O}(\exp(|\mathcal{X}|))$ entries. Hence, if we consider even small binary images whose pixel space is $\{0,1\}^{10\times 10}$, we would already need more evaluations than the number of atoms in the observable universe.

C2 **Many translations exist, but some are not sound.** As we will discuss later, not all translations lead to commutative diagrams. Therefore, it is important to consider how to construct such a translation.

C3 **Many models $P_{Y|X}$ might exist, but some may not satisfy desirable human properties.** Hence, simply verifying inference equivariance does not mean that users understand the internal behaviour of $P_{Y|X}$.

Conclusion 1. A user's ability to predict a model's outputs is essential (but not enough) for the model $P_{Y|X}$ to be interpretable.

These questions highlight the limits of informal definitions in the literature, that is, without a formal framework, key questions are more easily overlooked. In the following sections, we address each question by introducing a corresponding symmetry.

2.2. Symmetry II: Information Invariance

We begin by addressing challenge **C1**, namely the intractability of verifying inference equivariance for all possible inputs when the input space is large. Our first goal will then be to answer the following question: **(RQ2)** *How can we make inference equivariance tractable?*

Ideally, one would like to work with a model $P_{Y|X}$ for which inference equivariance can be verified in only a few steps. When this is not possible, we need to rewrite $P_{Y|X}$ to make inference equivariance more tractable. This requires a form of compression which captures *only the essential properties* of each input $x \in \mathcal{X}$ related to Y. To this end, we introduce a surjective map $P_{Z|X} : \mathcal{X} \twoheadrightarrow \mathbb{P}(Z)$ that, when composed with a model $P_{Y|Z}$, enables us to obtain $P_{Y|X}$ again. The effect of this transformation can be characterised by analysing the information content in Z compared to X (quantified by entropy $H : V \to \mathbb{R}$) and how much information we have left in Z to predict Y compared to X (quantified by the mutual information $I : V_1 \times V_2 \to \mathbb{R}$). Specifically, we argue that **C1** can be circumvented if the surjection $P_{Z|X}$ satisfies the following invariance:

Symmetry 2. **(Information Invariance)** Given a model $P_{Y|X}$, the mutual information $I(Y;X)$ is invariant under marginalisation $P_{Y|Z} \circ P_{Z|X}$ with $H(Z) \ll H(X)$ if the following diagram commutes:

$$\begin{tikzcd} X \arrow[r, two heads, "P_{Z|X}"] \arrow[dr, "P_{Y|X}"'] & Z \arrow[d, "P_{Y|Z}"] \\ & Y \end{tikzcd}$$

This invariance guarantees that Z retains all information in X that is relevant for predicting Y while discarding extraneous variability.

Remark 2.1. The marginalisation $P_{Y|Z} \circ P_{Z|X}$ makes the variable Y **conditionally independent** with respect to X given Z, that is, $I(Y;X \mid Z) = 0$.

If Y is conditionally independent with respect to X, then X does not provide any information to explain Y once we know Z. So, when verifying inference equivariance for a model $P_{Y|Z}$, we can safely ignore X. Since $H(Z) \ll H(X)$ by construction, the conditional probability table representing $P_{Y|Z}$ is exponentially smaller than $P_{Y|X}$. This leads to the following conclusion:

Conclusion 2. Verifying inference equivariance for $P_{Y|Z}$ is equivalent but more tractable than for $P_{Y|X}$.

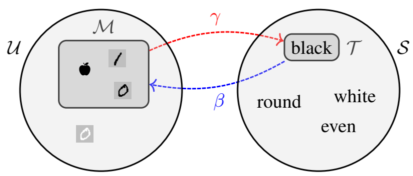
Consequences of Information Invariance ***Derivable properties.*** Sparsity (few features, few parameters), compactness (exclusion of irrelevant information) (Murphy, 2023), completeness (explanations are sufficient statistics of the model prediction) (Yeh et al., 2020), and modularity (a model can be broken down into simpler components) (Murdoch et al., 2019) could be derived from information invariance rather than being primitive notions. ***Methods to guarantee information invariance.*** Models employing feature selection (Miller, 1984), such as sparse decision trees, or leveraging the manifold hypothesis (Bengio et al., 2013), such as deep neural networks (Tishby & Zaslavsky, 2015), can enforce information invariance by design. ***Disqualified approaches.*** Traditional feature-attribution (Ribeiro et al., 2016; Lundberg & Lee, 2017; Erhan et al., 2009; Sundararajan et al., 2017) operates in the original input space and does not guarantee the existence of a lower-dimensional representation Z that captures all and only the information relevant for Y.

2.3. Symmetry III: Concept Closure Invariance

Symmetry 2 allows us to consider inference equivariance exclusively on $P_{Y|Z}$, but, as framed by challenge **C2**, it does not give us a criterion for characterising translations that enable the inference diagram to commute. Therefore, we now study which properties make a translation *sound*, that is, a translation that preserves "meaning" In other words, we now explore the following research question: **(RQ3)** *What is required for a translation to be sound?*

To address this question, we first introduce a key data structure —namely, *concepts*—which will allow us to characterise sound translations. Following Goguen (2005); Ganter & Wille (1996), we think of a **concept** as a relation between a set of objects (e.g., $\{$, , , $\ldots\}$) and symbols (e.g., $\{\texttt{black}, \texttt{round}, \ldots\}$) as shown in the following example:

Example 2.2. Consider a set of sentences $\mathcal{S} = \{\texttt{black}, \texttt{white}, \texttt{round}, \texttt{even}\}$ and a set of objects $\mathcal{U}$. For instance, we can fully define what we mean with the concept "black" via the tuple $(\mathcal{T}, \mathcal{M})$ with $\mathcal{T} = \{\texttt{black}\} \subseteq \mathcal{S}$ and $\mathcal{M} = \{$, , $\} \subseteq \mathcal{U}$, as shown in the next figure:



Given a set of objects $\mathcal{U}$ and a set of sentences $\mathcal{S}$, to define formally a concept we consider the functions $\beta : \mathcal{P}(\mathcal{S}) \to \mathcal{P}(\mathcal{U})$, which maps any $\mathcal{T} \subseteq \mathcal{S}$ into the subset of objects of $\mathcal{U}$ whose elements satisfy every sentence φ in $\mathcal{T}$, and $\gamma : \mathcal{P}(\mathcal{U}) \to \mathcal{P}(\mathcal{S})$, which maps any $\mathcal{M} \subseteq \mathcal{U}$ into the subset of all sentences in $\mathcal{S}$ satisfied by every object in $\mathcal{M}$. In summary a *concept* can be defined as follows:

Definition 2.3 (Concept)**.** Given a set of objects $\mathcal{U}$, a set of sentences $\mathcal{S}$, and functions β, γ as above, a **concept** is a tuple $(\mathcal{T}, \mathcal{M}) \in \mathcal{P}(\mathcal{S}) \times \mathcal{P}(\mathcal{U})$ such that the following diagram commutes

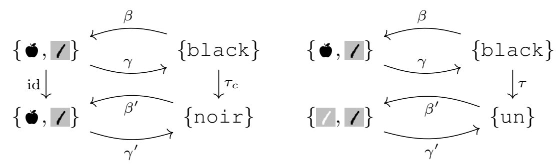
$$\begin{tikzcd} \mathcal{M} \arrow[r, bend right, "\gamma"'] & \mathcal{T} \arrow[l, bend right, "\beta"'] \end{tikzcd}$$

Under this definition, we will characterize a *sound translation* as a *concept-preserving* map associating different symbols (e.g., $\texttt{black}$ and $\texttt{noir}$) to the same objects (e.g., $\{$, , , $\ldots\}$). Intuitively, concepts allow the characterisation of sound translations, that is, translations that "preserve concepts" by noticing that, if an object satisfies a sentence φ, it should also satisfy the translated sentence $\tau(\varphi)$. We refer to such sound translations as *concept-based translations* τ_c. Concept-based translations are sound, and therefore enable interpretability via inference equivariance, if they satisfy the following invariance:

Symmetry 3. **(Concept Closure Invariance)** Concept closure is invariant under a sentence translation function $\tau_c : \mathcal{T} \to \mathcal{T}'$ if, for any pair of concepts $C = (\mathcal{T}, \mathcal{M})$ and $C' = (\mathcal{T}', \mathcal{M}')$, the following diagram commutes for all objects $\omega \in \mathcal{M}$:

$$\begin{tikzcd} \mathcal{M} \arrow[r, bend right, "\gamma"'] \arrow[d, "\mathrm{id}"'] & \mathcal{T} \arrow[l, bend right, "\beta"'] \arrow[d, "\tau_c"] \\ \mathcal{M}' \arrow[r, bend right, "\gamma'"'] & \mathcal{T}' \arrow[l, bend right, "\beta'"'] \end{tikzcd}$$

Example 2.4. Given a set of objects $\mathcal{U} = \{$, , $\}$, the sentences $\mathcal{T} = \{\texttt{black}\}$ and $\mathcal{T}' = \{\texttt{noir}, \texttt{un}\}$, and the objects $\mathcal{M} = \{$, $\} \subset \mathcal{U}$, the translation $\tau_c = \{\texttt{black} \to \texttt{noir}\}$ is sound as it preserves concept closure, while $\tau = \{\texttt{black} \to \texttt{un}\}$ is not



Having defined concepts and sound translations, we now give a probabilistic interpretation of concepts in order to characterise concept-based models for which these translations are sound.

Probabilistic interpretation of concepts To reason about concepts in a probabilistic setting, for each concept indexed by i, we denote as $\mathcal{M}_i$ the set of objects associated with that concept. We then introduce a map $g_i : \mathcal{X} \to \mathcal{C}$, such that, for each representation $x \in \mathcal{X}$, the value $g_i(x)$ represents possible concept outcomes in $\mathcal{C}$ associated with the i-th concept. We can then define a concept random variable as a map $C_i : \Omega \to \mathcal{C}$ defined by the composition:

$$\begin{tikzcd} \Omega \arrow[r, "X"] \arrow[dr, "C_i"'] & \mathcal{X} \arrow[d, "g_i"] \\ & \mathcal{C} \end{tikzcd}$$

A concept-based model can then be obtained by marginalising $P_{Y|Z}$ as $P_{Y|C} \circ P_{C|Z}$ where $P_{C|Z} : \mathcal{Z} \to \mathbb{P}(C)$ is a *concept encoder* and $P_{Y|C} : \mathcal{C} \to \mathbb{P}(Y)$ is a *concept-based task predictor*. This marginalisation makes Y conditionally independent to Z given C, which leads us to conclude that:

Conclusion 3. Verifying inference equivariance for $P_{Y|C}$ is equivalent to that for $P_{Y|Z}$, but it supports the use of sound translations.

Consequences of Concept-Closure Invariance ***Derivable properties.*** Alignment with a user's vocabulary (Kim et al., 2018) is formally expressed by this invariance. Faithfulness and fidelity (Jacovi & Goldberg, 2020) emerge as derived properties, since a translation that is not closed over the target concept space may induce multiple, potentially conflicting interpretations of the same model behaviour. ***Methods to guarantee concept-closure invariance.*** *Concept Bottleneck Models* (CBMs) (Koh et al., 2020) enforce concept-closure by construction by forcing the model to use concepts from the user's conceptual space. ***Disqualified approaches.*** Traditional machine learning models, such as decision trees (Breiman et al., 1984; Hu et al., 2019) or additive models (Hastie & Tibshirani, 1986; Agarwal et al., 2021), are disqualified whenever they are applied to non-concept spaces, such as pixel intensities.

2.4. Symmetry IV: Structural Invariance

Symmetry 2 and Symmetry 3 reduce the analysis of inference equivariance to $P_{Y|C}$. However, the "behaviour" of this model has not yet been characterised. To do this, we consider challenge **C3** and ask the following research question: ***(RQ4)*** *How should an interpretable model behave?*

The "behaviour" of a model is determined by its *structural properties* (such as linearity or monotonicity) which describe how outputs are computed from inputs. Knowing these properties allows a user to know what the model can and cannot do, enabling reasoning and control over its behaviour. Note that, which properties matter is user dependent, as they define the user's hypothesis space Hm of "mental models". When $P_{Y|C}$ satisfies the structural properties encoded in Hm, the user can internally simulate the model, as required by inference equivariance.

Example 2.5. A student hypothesis space might be the space of linear models $\mathbb{E}[Y \mid C] = wC$, which might be different from a researcher hypothesis space including quadratic models $\mathbb{E}[Y \mid C] = w_1 C + w_2 C^2$. If $P_{Y|C}$ takes a quadratic form, then the researcher could understand the model's behaviour, while the student will not.

To formalise this intuition, we introduce a functor F between a category of models Im (standing for interpretable models) and a category of models Hm (standing for human mental models). Structural invariants are then defined as the properties preserved under F, that is, the properties that both Im and Hm models have in common.

Symmetry 4. **(Structural Invariance)** The structural properties of models in Im are invariant under the functor $F : \mathsf{Im} \to \mathsf{Hm}$ if and only if there exist two injective functors $E_1 : \mathsf{Im} \to \mathsf{Stoch}$ and $E_2 : \mathsf{Hm} \to \mathsf{Stoch}$ such that the following diagram commutes:

$$\begin{tikzcd} \mathsf{Im} \arrow[r, hook, "E_1"] \arrow[d, "F"'] & \mathsf{Stoch} \\ \mathsf{Hm} \arrow[ur, "E_2"'] & \end{tikzcd}$$

Conclusion 4. If $P_{Y|C}$ is compatible with the hypothesis space of user's mental models, then the user can internally simulate the model to verify inference equivariance.

Consequences of Structural Invariance ***Derivable properties.*** Structural invariance makes interpretability explicitly user-centric and task-specific: the structural properties required of a model depend on the user's "mental model" (Johnson-Laird, 1983; Kulesza et al., 2013; Kim et al., 2018). These properties yield translucence as it exposes the model's specific hypothesis spaces. As a result, properties such as linearity, monotonocity, or sparsity are not first principles but instantiations of structural properties chosen to match a given user. ***Methods to guarantee structural invariance.*** The most common approach is to fix the hypothesis space of the model a priori, for instance by restricting to linear models, or monotone functions (Rudin, 2019; Debot et al., 2024). More flexibility can be achieved with models that explicitly factorise concept dependencies (Dominici et al., 2024). ***Disqualified approaches.*** Sparse autoencoders and concept-based models whose mappings between concepts and tasks (or between concepts themselves) are arbitrary, such as employing DNNs, have hypothesis space that does not satisfy any interesting structure, and therefore do not satisfy this symmetry. More broadly, models that violate domain-required structures must be excluded: for example, in medical risk scoring, where monotonicity is a critical structural requirement, highly expressive models such as unconstrained deep neural networks are not interpretable for that domain, regardless of their predictive accuracy.

3. Probabilistic Interpretable Models

In this section we leverage the discussed symmetries to formally answer the following question: ***(RQ5)*** *How to formalise interpretable models?*

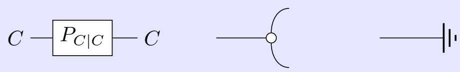
We will answer this question by first giving a complete formal definition of an *interpretable model* in terms of its symmetries, and then presenting the *category of interpretable models*, that is, a general "recipe" to build interpretable models.

Definition 1 (Interpretable Models). Interpretable models are those that satisfy all interpretability symmetries.

The general category of interpretable models (as a subcategory of probabilistic models), is specified by the objects, processes, and compositional rules from which any interpretable model can be constructed.

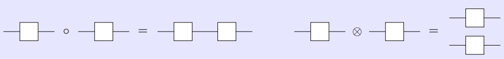
Definition 2. (Category of Interpretable Models) The category of interpretable models $\mathsf{Im} \subseteq \mathsf{Stoch}$ has:

- **Processes**: concept-based conditional probability distributions $P_{C|C}$, copy maps (duplicating a variable), and discard maps (marginalising a variable).



Concept-based model   Copy Map   Discard Map

- **Composition rules**: sequential and parallel composition of processes.



Sequential   Parallel

Remark 3.1. Note that, if Y is a concept-based random variable, then $P_{Y|C}$ is a valid process in this category. Moreover, we can enrich this category with processes involving X, Z, and other variables to build arbitrary complex models.

Read from left to right, diagrams constructed using the above rules are known as *string diagrams* (Baez & Stay, 2010). These diagrams form a formal visual language for representing probabilistic models (Jacobs et al., 2025). Their key advantage is that they also represent probabilistic inference diagrammatically, providing a unified way to reason about inference in interpretable models.

4. Inference on Probabilistic Interpretable Models

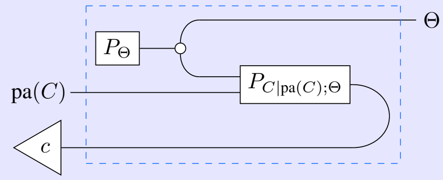
Now that we formalised interpretable models, this section describes inferences on such models, that is, how interpretable models can be used in practice. We will describe three of the most common and widely used types of inference used in interpretability: concept alignment (Sec 4.1), interventions, and counterfactuals (Sec 4.2).

4.1. Concept Alignment

In this section, we discuss how the inference process allows the model to align its concepts w.r.t. a target user. We therefore ask the following question: ***(RQ6)*** *How to learn human concepts?*

The process of matching a model's concepts to human concepts is called *alignment*. In a nutshell, we aim to compute the distribution of parameters Θ that make the prediction of a concept map match a set of ground-truth concepts c. From a probabilistic perspective, this corresponds to finding the posterior of the parameters Θ considering as evidence a set of ground-truth concepts c. In diagrammatic terms, we can represent observing the evidence $C = c$ by bending the C wire backwards and inserting a constant state representing the evidence c. We indicate the normalisation integral in the Bayesian posterior with a blue dashed box and with $\text{pa}(C)$ the inputs of a model with output C.

Definition 3 (Concept Alignment). Given a parametric concept map $P_{C|\text{pa}(C);\Theta}$, concept alignment is the Bayesian inversion represented by the diagram:



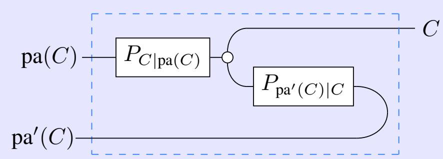
$$P_{\Theta|C=c,\text{pa}(C)}(\theta) = \frac{P_{C|\text{pa}(C);\Theta}(c|\text{pa}(C);\theta)P_\Theta(\theta)}{\int_\Theta P_{C|\text{pa}(C);\Theta}(c|\text{pa}(C);\theta)P_\Theta(\theta)\mathrm{d}\theta}$$

4.2. Human-machine interaction

Given an aligned interpretable model, we can now turn our attention to how humans may interact with it. In particular, we ask: ***(RQ7)*** *Which queries are supported by interpretable models?*

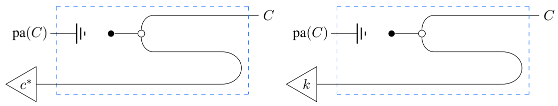
Interventional inference A key advantage of concept-based models is their support for human interaction, allowing users to *intervene* and change concept activations in various ways (Koh et al., 2020; Chauhan et al., 2022; Barker et al., 2023; Shin et al., 2023; Collins et al., 2023; Espinosa Zarlenga et al., 2023; Marcinkevičs et al., 2024). In general interventions can be described as a form of posterior inference where we consider a given concept-based process $P_{C|\text{pa}(C)}$ as a *prior* and introduce a *likelihood* $P_{\text{pa}'(C)|C}$ to update our prior by observing as evidence other random variables $\text{pa}'(C)$.

Definition 4 (Intervention). Given a parametric concept map $P_{C|\text{pa}(C)}$ and a likelihood $P_{\text{pa}'(C)|C}$, an intervention is the Bayesian inversion represented by the diagram:



$$P_{C|\text{pa}(C),\text{pa}'(C)}(c \mid \text{pa}(c), \text{pa}'(c)) = \frac{P_{\text{pa}'(C)|C}(\text{pa}'(c)|c)\, P_{C|\text{pa}(C)}(c|\text{pa}(c))}{\int_C P_{\text{pa}'(C)|C}(\text{pa}'(c)|c)\, P_{C|\text{pa}(C)}(c|\text{pa}(c))\,\mathrm{d}c}$$

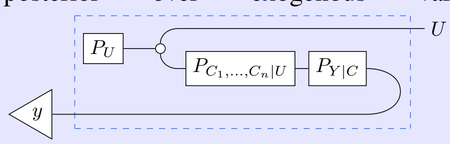
Under these lenses, ground-truth interventions (left image) $P_{C|gt(\text{pa}(C),\text{pa}'(C))} = P_{C|\text{pa}'(C)=c^*}$ become a special case where the prior becomes a uniform distribution independent on $\text{pa}(C)$ (which gets discarded), the likelihood is the identity function, and the evidence is represented by ground-truth labels c^*, while do-interventions (right image) $P_{C|do(\text{pa}(C),\text{pa}'(C))} = P_{C|\text{pa}'(C)=k}$ are a special case where the evidence is represented by a constant value $k \in \mathbb{R}$:



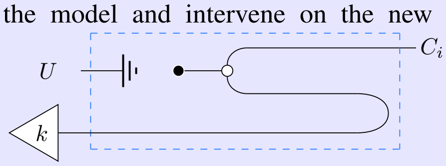
Counterfactual inference Counterfactual inference is a cornerstone of both interpretable and causal machine learning, as it enables the evaluation of causal effects in hypothetical scenarios. Following the standard causality approach (Jacobs et al., 2025), counterfactual inference requires rewriting a given model in such a way that all randomness is confined in a set of source distributions P_{U_i} referred to as "exogenous" variables that capture high-dimensional and potentially entangled information.

Definition 5 (Counterfactual). Given a model $P_{Y|C_1,\dots,C_n} \circ P_{C_1,\dots,C_n|U} \circ P_U$ a counterfactual is the following sequence of Bayesian inversions:

- **Abduction:** observe the value of certain concepts Y to compute the posterior over exogenous variables:



- **Action:** given the posterior on U, duplicate the model and intervene on the new model:



- **Prediction:** infer the value of Y in the duplicated intervened model $P_{Y|do(C)}$.

Conclusion 5. Alignment, interventional, and counterfactual inference are all forms of Bayesian inversion.

Consequences of Probabilistic Interpretability ***Derivable properties.*** Alignment, interventions, and counterfactuals have actionability (Poyiadzi et al., 2020) and interactivity (Tenney et al., 2020) as direct consequences. Moreover, alignment can be formally measured within this framework, inheriting the same theoretical guarantees and generalisation bounds as probabilistic inference itself. ***Methods for probabilistic interpretable inference.*** Unifying alignment, interventional, and counterfactual inference means that a single posterior inference algorithm, such as belief propagation, can, in principle, be employed to perform all interpretable inferences, providing a coherent and computationally grounded foundation for probabilistic interpretable inference.

5. Alternative Views

Alternative position: there is no universal, mathematical definition of interpretability, and there never will be. As discussed, most works in the interpretability literature define their subject matter in informal terms. As a consequence, the literature has gravitated toward enumerating extensive lists of desirable properties (modularity, simplicity, stability, completeness, actionability, etc.) rather than converging on unifying principles. Moreover, many authors observed how interpretability is necessarily user-centric and task-specific. This fragmentation and relativism has led Murphy (2023) to conclude that *there is no universal, mathematical definition of interpretability, and there never will be*. **Rebuttal:** While this view reflects the current state of the field, the absence of formal grounding has generated significant ambiguity, undermining the coherence and progress of interpretable AI research. Our position resolves this fragmentation by identifying a small set of general formal principles from which the properties discussed in the literature can be derived. Importantly, Symmetries I, III, and IV explicitly incorporate a human reference, thereby accounting for user-centrism and task-dependence without abandoning mathematical universality.

Alternative position: symmetries are not enough to define interpretability A potential objection is that symmetries alone may be insufficient to capture all essential aspects of interpretability. In particular, several properties commonly associated with interpretability do not have yet an agreed-upon formalisation in the literature (e.g., there are many different perspectives on what "simple" is), making it difficult at present to establish whether symmetries are adequate to account for them. Moreover, interpretable AI remains a relatively young field, and future research may uncover additional properties that require either further symmetries or an altogether different formal framework. **Rebuttal:** The introduced symmetries are designed to capture existing properties that have already been discussed in the literature, and we explicitly indicate how specific interpretability properties might be derived as consequences of specific symmetries. In this sense, the framework is extensible: newly identified properties can be analysed in terms of whether they correspond to additional symmetries or motivate refinements of the existing ones.

Alternative position: symmetries are not suitable for the interpretability community Even if symmetries are expressive enough in principle, a further concern is that the proposed symmetries are formulated using mathematical notions that may be unfamiliar to many researchers in the interpretability community. This could create a barrier to understanding and adoption, undermining their practical value as a foundation for interpretability. **Rebuttal:** There is an inherent trade-off between formal rigour, which enables clarity and generality, and accessibility, which lowers barriers to entry. Our formalisation aims to strike a balance between these competing goals: while mathematically precise, each symmetry admits intuitive, example-level interpretations that can be understood without full mastery of the underlying formalism. This allows a broad audience to develop an operational intuition for the proposed first principles, without sacrificing the generality and unifying power that formalisation affords.

6. Call to Action

To realise the aims of this position, the community should prioritise a systematic formal validation of the proposed first principles by showing that existing interpretability desiderata and evaluation criteria can be derived as consequences of the proposed symmetries. This requires researchers explicitly stating which symmetry assumptions their methods rely on and which interpretability properties follow from them, thereby clarifying when a method is interpretability-complete with respect to the above symmetry and when it is not.

As a concrete example, the concept-based interpretability community (Poeta et al., 2023) has largely focused on concept-closure invariance, but has paid comparatively little attention to structural invariance. As a result, concept-based models often exhibit arbitrary behaviours (e.g., when using DNNs for task predictors) or are weak classifiers (e.g., when using linear models for task predictors). Conversely, the mechanistic interpretability community has been less attentive to concept-closure invariance and information invariance. This typically leads to concepts that are not aligned with human semantics and to information overload.

We therefore call on researchers to (i) map existing interpretability methods to the proposed symmetries, (ii) identify which symmetries are implicitly assumed or violated, and (iii) develop new methods that are explicitly symmetry-complete with respect to the interpretability goals they aim to satisfy.

References

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., and Hinton, G. E. Neural additive models: Interpretable machine learning with neural nets. *Advances in neural information processing systems*, 34: 4699–4711, 2021.

Alvarez-Melis, D. and Jaakkola, T. S. Towards robust interpretability with self-explaining neural networks. *arXiv preprint arXiv:1806.07538*, 2018.

Baez, J. and Stay, M. Physics, topology, logic and computation: a rosetta stone. In *New structures for physics*, pp. 95–172. Springer, 2010.

Barker, M., Collins, K. M., Dvijotham, K., Weller, A., and Bhatt, U. Selective concept models: Permitting stakeholder customisation at test-time. *AAAI HCOMP*, 2023.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Biran, O. and Cotton, C. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pp. 8–13, 2017.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. Classification and regression trees. 1984.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Chauhan, K., Tiwari, R., Freyberg, J., Shenoy, P., and Dvijotham, K. Interactive concept bottleneck models. *arXiv preprint arXiv:2212.07430*, 2022.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019.

Collins, K. M., Barker, M., Espinosa Zarlenga, M., Raman, N., Bhatt, U., Jamnik, M., Sucholutsky, I., Weller, A., and Dvijotham, K. Human uncertainty in concept-based ai systems. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 869–889, 2023.

Colombini, J. J., Bonchi, F., Giannini, F., Giannotti, F., Pellungrini, R., and Frosini, P. Mathematical foundation of interpretable equivariant surrogate models. In *World Conference on Explainable Artificial Intelligence*, pp. 294–318. Springer, 2025.

Debot, D., Barbiero, P., Giannini, F., Ciravegna, G., Diligenti, M., and Marra, G. Interpretable concept-based memory reasoning. *Advances in Neural Information Processing Systems*, 37:19254–19287, 2024.

Dominici, G., Barbiero, P., Espinosa Zarlenga, M., Termine, A., Gjoreski, M., Marra, G., and Langheinrich, M. Causal concept graph models: Beyond causal opacity in deep learning. *arXiv preprint arXiv:2405.16507*, 2024.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Erhan, D., Bengio, Y., Courville, A., and Vincent, P. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

Espinosa Zarlenga, M., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lio, P., and Jamnik, M. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in Neural Information Processing Systems*, 35:21400–21413, 2022.

Espinosa Zarlenga, M., Collins, K., Dvijotham, K., Weller, A., Shams, Z., and Jamnik, M. Learning to receive help: Intervention-aware concept embedding models. *Advances in Neural Information Processing Systems*, 36:37849–37875, 2023.

Facchini, A. and Termine, A. A first contextual taxonomy for the opacity of ai systems. 2021.

Fioravanti, S., Giannini, F., Barbiero, P., Frazzetto, P., Confalonieri, R., Zanasi, F., and Navarin, N. Categorical explaining functors: Ensuring coherence in logical explanations. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 22, pp. 306–315, 2025.

Fritz, T. A synthetic approach to markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 370:107239, 2020.

Ganter, B. and Wille, R. Formal concept analysis. *Wissenschaftliche Zeitschrift-Technischen Universitat Dresden*, 45:8–13, 1996.

Giannini, F., Fioravanti, S., Barbiero, P., Tonda, A., Liò, P., and Di Lavore, E. Categorical foundation of explainable ai: A unifying theory. In *World Conference on Explainable Artificial Intelligence*, pp. 185–206. Springer, 2024.

Goguen, J. What is a concept? In *International Conference on Conceptual Structures*, pp. 52–77. Springer, 2005.

Hastie, T. and Tibshirani, R. Generalized additive models. *Statistical science*, 1(3):297–310, 1986.

Hu, X., Rudin, C., and Seltzer, M. Optimal sparse decision trees. *Advances in neural information processing systems*, 32, 2019.

Jacobs, B., Széles, M., and Stein, D. Compositional inference for bayesian networks and causality. *Math. Found. of Programming Semantics*, 2025.

Jacovi, A. and Goldberg, Y. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*, 2020.

Johnson-Laird, P. N. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge University Press, Cambridge, UK, 1983. ISBN 9780521273912.

Kim, B., Khanna, R., and Koyejo, O. O. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Klein, F. A comparative review of recent researches in geometry. *Bulletin of the American Mathematical Society*, 2(10):215–249, 1893.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., and Wong, W.-K. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*, pp. 3–10. IEEE, 2013.

Lee, C. K., Samad, M., Hofer, I., Cannesson, M., and Baldi, P. Development and validation of an interpretable neural network for prediction of postoperative in-hospital mortality. *NPJ digital medicine*, 4(1):8, 2021.

Lipton, Z. C. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

Lou, Y., Caruana, R., and Gehrke, J. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 150–158, 2012.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Marcinkevičs, R., Laguna, S., Vandenhirtz, M., and Vogt, J. E. Beyond concept bottleneck models: How to make black boxes intervenable? *arXiv preprint arXiv:2401.13544*, 2024.

Marconato, E., Passerini, A., and Teso, S. Interpretability is in the mind of the beholder: A causal framework for human-interpretable representation learning. *Entropy*, 25 (12):1574, 2023.

Meng, C., Trinh, L., Xu, N., Enouen, J., and Liu, Y. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1): 7166, 2022.

Miller, A. J. Selection of subsets of regression variables. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 147(3):389–410, 1984.

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

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*, 2019.

Murphy, K. P. *Probabilistic machine learning: Advanced topics*. MIT press, 2023.

Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T., and Baralis, E. Concept-based explainable artificial intelligence: A survey. *arXiv preprint arXiv:2312.12936*, 2023.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. Face: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350, 2020.

Ribeiro, M. T., Singh, S., and 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*, pp. 1135–1144, 2016.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Shin, S., Jo, Y., Ahn, S., and Lee, N. A closer look at the intervention procedure of concept bottleneck models. *arXiv preprint arXiv:2302.14260*, 2023.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Tenney, I., Wexler, J., Bastings, J., Bolukbasi, T., Coenen, A., Gehrmann, S., Jiang, E., Pushkarna, M., Radebaugh, C., Reif, E., et al. The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. *arXiv preprint arXiv:2008.05122*, 2020.

Tishby, N. and Zaslavsky, N. Deep learning and the information bottleneck principle. In *2015 IEEE information theory workshop (ITW)*, pp. 1–5. IEEE, 2015.

Tull, S., Lorenz, R., Clark, S., Khan, I., and Coecke, B. Towards compositional interpretability for xai. *arXiv preprint arXiv:2406.17583*, 2024.

Watson, D. S. and Floridi, L. The explanation game: a formal framework for interpretable machine learning. In *Ethics, governance, and policies in artificial intelligence*, pp. 185–219. Springer, 2021.

Yeh, C.-K., Kim, B., Arik, S., Li, C.-L., Pfister, T., and Ravikumar, P. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33:20554–20565, 2020.