Causal Articulation Theory (CAT): Articulating Static and Temporal Causal Models Beyond LLM Rationalizations

Matej Zečević MATEJ.ZECEVIC@TU-DARMSTADT.DE
Department of Computer Science, TU Darmstadt, Germany

Devendra Singh Dhami D.S.DHAMI@TUE.NL
Department of Mathematics and Computer Science, TU Eindhoven, Netherlands

Kristian Kersting KERSTING@CS.TU-DARMSTADT.DE
Department of Computer Science, TU Darmstadt;
and Hessian Center for AI (hessian.AI), and German Research Center for AI (DFKI), Germany

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

In normative terms, a meaningful explanation should reflect how data is generated—this is precisely where causality becomes essential. Existing machine learning methods for explanation either miss this desideratum entirely or incorporate only partial causal knowledge. Large language models (LLMs), in particular, routinely produce fluent explanations that amount to unfaithful rationalizations of the underlying data-generating process. We therefore advocate a direct approach: derive explanations from the structural parameters and mechanisms of a causal model itself. To this end, we introduce Causal Articulation Theory (CAT), a formal account of how explanations can be articulated from a structural causal model (SCM). CAT addresses why-questions about individual units and uses a recursive articulation procedure that draws on both the graphical structure and causal effects encoded in the SCM. Using a small set of first-order articulation rules, we first develop CAT for static linear SCMs and show that it offers an appealing alternative to LLM rationalizations: CAT naturally distinguishes direct from indirect causes, captures the qualitative sign of causal effects, and remains robust to stochasticity in exogenous variables. To address ever-changing environments, we then extend CAT beyond static settings by relaxing initial assumptions to cover both temporal and agentic scenarios, yielding articulated explanations over time and action. For empirical corroboration, we present a series of experiments: (i) a user study examining the alignment between CAT-based articulations and human causal judgments in everyday domains, (ii) an investigation of CAT as a regularizer for causal discovery, and (iii) examples of articulated explanations in two temporal domains involving forecasting and a simple video game environment.

Keywords: causal explanation, structural causal model, interactive machine learning, large language model

1. Introduction

Artificial intelligence research has long pursued the goal of automating aspects of scientific reasoning (McCarthy, 1998; McCarthy and Hayes, 1981; Steinruecken et al., 2019). Because causal interactions sit at the core of human cognition (Penn and Povinelli, 2007), integrating causal structure into machine reasoning remains a central challenge (Schölkopf, 2022), even in light of mature formalisms for causal representation and inference (Pearl, 2009; Peters et al., 2017). One long-standing promise of causality for AI lies in the prospect of explanations: causality and explanation are widely regarded as inseparable (Josephson and Josephson, 1996). According to Miller (2019),

explanations involve two processes—first, a cognitive step that identifies the causes of an event (possibly relative to counterfactuals), and second, a communicative step in which these causes are conveyed to an explainee. The question then becomes: *what constitutes a suitable representational substrate for such explanations?*

Pearl and Mackenzie (2018) argue that counterfactual, symbolic causal reasoning is fundamental to human-level intelligence and to the way humans generate and communicate explanations. This perspective positions the *structural causal model* (SCM) as a natural representational basis for explanations. A body of cognitive science supports this view, showing that SCM-like structures capture essential aspects of human causal reasoning (Gerstenberg et al., 2015, 2017) and therefore aspects of how humans explain (Lagnado et al., 2013). Moreover, humans—children included—routinely rely on counterfactual and why-questions to learn, explore, and navigate the world (Gopnik, 2012; Buchsbaum et al., 2012; Byrne, 2016). These patterns of reasoning are components of the human *mental model*, the internal representation of how the world works (Simon, 1961; Nersessian, 1992; Chakraborti et al., 2017). In this work, we adopt the common view that explanations answer *why-questions* (Dennett, 1989), meaning they are inherently counterfactual and therefore inherently causal.

Building on these insights, we introduce *Causal Articulation Theory* (CAT), a conceptual and procedural account of how explanations can be articulated directly from the mechanisms and parameters of an SCM. CAT defines the class of why-questions it addresses and provides a recursive articulation procedure that interprets causal structure—edge relations, signs, causal effects, and parameter magnitudes—to produce human-readable explanations grounded in the model. Unlike post-hoc explanation methods that analyze downstream predictors or selectively incorporate causal information, CAT treats the SCM itself as the explanatory substrate. This stands in contrast to large language models (LLMs), which generate fluent but unconstrained rationalizations of model behavior and therefore often fail to reflect true data-generating mechanisms. We stress that LLMs serve here purely as motivating contrast: CAT does not assume LLM-style inputs or outputs, nor does it require LLMs at any stage of the pipeline. Rather, CAT is a formal alternative to post-hoc rationalization, operating directly on structural causal mechanisms.

A motivating example illustrates the idea. The question "Why is the temperature at the Matterhorn low?" can be articulated by CAT as "Because the altitude is high," reflecting a largely time-independent physical relation. A more involved example is the classical "Causal Hans" scenario involving the variables *Age*, *Nutrition*, *Health*, and *Mobility*. CAT articulates the why-question "Why is Hans's mobility below average?" as "Hans's mobility is low because his health is poor, which is largely due to his high age, although his nutrition is good." While this articulation captures the essential causal pathways, many domains are not static: mobility may depend not only on current health but also on past mobility; past health may influence current nutrition; and so forth.

Real-world environments are fundamentally temporal. They often involve sequences of evolving states, persistent causal influences, and in many cases agents whose actions alter the data-generating process. To address this, we extend CAT to handle temporal and agentic causal models, enabling articulated explanations of outcomes that unfold over time or depend on action-conditioned dynamics. This extension introduces explanation structures that reflect both variable-to-variable causal relations and their propagation through time steps. Figure 9, for example, illustrates how CAT articulations reveal behavioral differences between two agents in a simple grid-world setting.

Contributions. Our contributions unify and extend two previously separate lines of work into a single conceptual and operational account:

(i) **Causal Articulation Theory (CAT).** We introduce CAT, a new account of causal explanation that articulates why-questions directly from the structural equations and parameters of an SCM, drawing on both graphical structure and model-implied causal effects.

(ii) **Human alignment.** We conduct a user study to examine how CAT articulations align with human causal judgments across everyday scenarios.

(iii) **CAT with learned causal models.** We feed CAT with causal representations learned from data and show how CAT reveals characteristic errors of causal discovery methods. We also propose a simple CAT-derived regularization penalty to reduce false edges in learned graphs.

(iv) **Temporal and agentic extension.** We extend CAT to cover time-dependent and agentic settings, introducing articulation rules and structures that handle temporal recurrences and action-conditioned causal mechanisms.

(v) **Empirical validation in temporal domains.** We evaluate temporal CAT on synthetic time-series data and a 2D grid-world game, demonstrating its ability to produce coherent, informative articulations of dynamic systems.

2. Prerequisites & Assumptions

We follow the formalism of (Pearl, 2009) for discussing causation but adapt a modern formalization inline with works such as (Bongers et al., 2021). The key input to our explainer is going to be an (approximation of an) SCM, which we define as:

Definition 1 (SCM) *A structural causal model is a tuple* $\mathcal{M} = (\mathbf{V}, \mathbf{U}, \mathbf{F}, \mathrm{P}_{\mathbf{U}})$ *forming a directed acyclic graph* $\mathcal{G}$ *over variables* $\mathbf{X} = \{X_1, \ldots, X_K\}$ *taking values in* $\boldsymbol{\mathcal{X}} = \prod_{k \in \{1 \ldots K\}} \mathcal{X}_k$ *subject to a strict partial order* $<_{\mathbf{X}}$*, where*

- $\mathbf{V} = \{X_1, \ldots, X_N\} \subseteq \mathbf{X}, N \leq K$ *is the set of endogenous variables.*
- $\mathbf{U} = \mathbf{X} \setminus \mathbf{V} = \{X_{N+1}, \ldots, X_K\}$ *is the set of exogenous variables.*
- $\mathbf{F}$ *is the set of deterministic structural equations,* $V_i := f_i(\mathbf{X}')$*, where* $\mathbf{X}' \subseteq \{X_j \in \mathbf{X} \mid X_j <_{\mathbf{X}} V_i\}$ *denoted by* $\mathrm{pa}(V_i)$ *are the parents of* V_i*.*
- $\mathrm{P}_{\mathbf{U}}$ *is the probability distribution over* $\mathbf{U}$*.*

For the causally curious reader we point to appendix Sec.C for discussions of our framework in the light of unobservable confounders and similar phenomena. Generally speaking, we start off with stronger assumptions while developing CAT from ground up and to ease the presentation to the first time reader and then gradually relax requirements when moving on to more complex domains. Therefore, we start off with a restricted class of SCMs: linear SCMs. Formally, we state:

Assumption 1 (Linear SCM) *The SCMs under consideration have linear structural equations, that is,* $\mathbf{F} \subset \{f \mid f(\mathrm{pa}(v)) = \boldsymbol{\alpha}^\top \mathrm{pa}(v), \boldsymbol{\alpha} \in \mathbb{R}^{|\mathrm{pa}(v)|}\}$.

We begin with linear SCMs to keep the articulation rules precise and interpretable. The framework itself does not fundamentally depend on linearity: CAT operates on the *sign and magnitude* of causal effects, and the temporal extension already relaxes several of these assumptions. Linear SCMs therefore serve as a transparent, foundational case rather than a restriction of principle. While the above assumption will be sufficient for the discussion of discrete random variables, our results naturally extend to continuous random variables w.l.o.g. if the following assumption on the exogenous variables holds:

Assumption 2 (Gaussian SCM) *The SCMs under consideration have normally distributed exogenous variables, that is,* $\mathrm{P}_{\mathbf{U}} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *with* $\boldsymbol{\mu} \in \mathbb{R}^{|\mathbf{U}|}, \boldsymbol{\Sigma} \in \mathbb{R}^{|\mathbf{U}|^2}$.

Since an explanation demands the discussion of causes, a very useful object for actually capturing the amount of influence a cause has on its effect is the *causal effect* which is defined as the interventional distribution as follows:

Definition 2 (CE) *For some target of interest* V_i *and a set of variables* $\mathbf{V}' \subset \mathbf{V}$ *an intervention* $do(\mathbf{V}' = \mathbf{v}')$ *replaces all original structural equations* $\{f_j\}_{V_j \in \mathbf{V}'}$ *by the constant assignment* $V_j' := v_j'$. *The induced distribution* $p(v_i \mid do(\mathbf{v}'))$ *is called causal effect of* $\mathbf{V}'$ *on* V_i.

Lastly, to quantify the causal effect for any given random variable pair in a single scalar value, we can resort to the average effect. We have that:

Definition 3 (ACE) *For a pair of random variables* $(V_i, V_j) \in \mathbf{V}$ *that satisfy*

(D) $\frac{1}{k-l}\big(\mathbb{E}[V_i \mid do(V_j = k)] - \mathbb{E}[V_i \mid do(V_j = l)]\big) = \alpha_j$ *with* $k, l \in \mathcal{X}_j$ *in the case that* (V_i, V_j) *are discrete random variables or*

(C) $\frac{\partial}{\partial v_j}\mathbb{E}[V_i \mid do(v_j)] = \alpha_j$ *where* (V_i, V_j) *are continuous*

we call α_j *the average causal effect (ACE) of* V_j *on* V_i. *For a set* $\mathbf{V}' \subset \mathbf{V}$ *of causes,* $\boldsymbol{\alpha} = (\alpha_j)_{V_j \in \mathbf{V}'}$ *collects all pairwise ACEs.*

Remember the notation of the expected value for a discrete random variable pair X, Y as $\mathbb{E}[Y \mid x] := \sum_y y \cdot p(y \mid x)$ and for continuous RVs as $\mathbb{E}[Y \mid x] := \int y \cdot p(y \mid x)\, dy$, where $\mathbb{E}[Y \mid do(X)]$ then refers to the expected causal effect, that is, the expected value of Y under intervention $do(X)$ replacing the conditional distribution $p(y \mid x)$ with the causal effect $p(y \mid do(x))$. In a last step, we connect our assumptions with the definition of ACE.

Observation 1 *In SCMs that follow Assumption 1 the coefficients* $\boldsymbol{\alpha}$ *of the structural equations are average causal effects since Def.3(D) is satisfied. For continuous random variables, Assumption 2 is sufficient to satisfy Def.3(C).*

This observation tells us that with the above definitions and assumptions when deriving our explanations we will be able to simply treat our SCM as a weighted adjacency matrix of the endogenous variables, which in turn will simplify computation immensely and make our explanations compatible with a wide range of existing graph learning algorithms.

3. Causal Articulation Theory First Introduction (Static, Linear SCM)

While acknowledging the difficulty of the problem and its philosophical nature, we address it pragmatically in a step-by-step derivation that leverages qualitative knowledge on SCMs as defined in the previous section. This connection will justify the naming as *Structural Causal* Explanation. Our running example, that of patient Hans, is a homage to the famous fallacy in explainable AI and psychology known as "Clever Hans" named after the 20th century Orlov Trotter horse Hans that was wrongly believed to be able to perform arithmetic (Pfungst, 1911). A "Clever Hans" moment is failure due to spurious associations in the data. For example, an image classifier that learns on watermarked images will have high accuracy on the test data from the same distribution by predicting the class using the watermark labels (that is, the model is "right for the wrong reasons") and furthermore fails completely when moving out-of-distribution (Lapuschkin et al., 2019). Some works such as (Stammer et al., 2021) moved beyond basic methods (like heat maps for image data) by employing expert intervention to move beyond such "Clever Hans" fallacies. Since explanations ought not only be "clever" but also causal, we will refer to our running example as the "Causal Hans" example. In the following, consider an SCM as before that generates medical records described by numerical representations for age, nutrition, overall health and mobility respectively ($\mathbf{V} = \{A, N, H, M\}$). Next, let's consider some samples from said SCM. E.g. we might observe the data set containing the individual named Hans $\mathbf{H} = (H_A, H_N, H_H, H_M) = (93.8, 58.8, 2.6, 26.2)$ where for sake of simplicity each value could be associated with a discrete label e.g. $H_A = 93.8$ would be 93 years old, whereas $H_M = 26.2$ could refer to a rather immobile person. The latter label is actually implicitly the assessment $H_M < \mu_M$ where $\mu_M = 35.6$ is the population's average mobility value, that is, we observe Hans to be a rather immobile person *when compared (or relative) to the other patients*. The population average values for our running example are $\boldsymbol{\mu} = (\mu_A, \mu_N, \mu_H, \mu_M) = (62.6, 32.8, 45.1, 35.6)$. With this we are in the position to pose a question like

Q1: *"Why is Hans's Mobility bad?"*

where the word "bad" refers to "bad relative to the population." Formally, we can now define such a question as:

Definition 4 (Why Question) *A quantity $Q_i := R(v_i, \mu_i)$ with binary ordering $R \in \{<, >\}$ where $V_i \in \mathbf{V}$ and μ_i is the empirical mean value for V_i, is called why-question concerning V_i if the ordering holds true, that is, $\mathbb{1}_R(Q_i) = 1$.*

Remember the notation for the indicator function $\mathbb{1}_R(Q_i) = 1$ if $(v_i, \mu_i) \in R$ and 0 otherwise. Checking back with the definition, we see that **Q1** defines a valid question for the Causal Hans example since $Q_M := H_M < \mu_M = 26.2 < 35.6$ holds true in our example data. On another note, we call Q_i a why-question because it relates to the *counterfactual* scenario regarding the causes of V_i, for example, how would've age, nutrition and health had to be if we were to think that Hans' mobility was not bad. While the number of valid why-questions that can be asked seems limited at first sight, the number scales linearly with the SCM as it is coupled to the endogenous variables. Specifically it is $\mathcal{O}(|\mathbf{V}|)$ thus we can potentially ask a question for any endogenous variable of an arbitrarily large SCM.

Next, we will discuss the knowledge on the SCM that our explanation will leverage. Generally, the true data-generating SCM $\mathcal{M}^*$ is unobserved but we can realistically expect to have access to an estimate of $\mathcal{M}^*$. Let's consider following SCM estimate $\mathcal{M}$ that contains the relations $A \xrightarrow{\alpha} N$,

$A \xrightarrow{\beta} H$, $N \xrightarrow{\gamma} H$, $H \xrightarrow{\delta} M$ where $\alpha, \beta, \gamma, \delta$ denote the respective (average) causal effects. Further, $\alpha, \gamma, \delta > 0$ while $\beta < 0$. That is, $\alpha > 0$ means that increasing age by a single unit increases nutrition by α units. Similarly, $\beta < 0$ means that for any unit increment of age we will have a β number of units decrement of health. Furthermore comparing between coefficients, $\beta > \gamma$ means that the causal effect of aging on health is greater in absolute terms than the causal effect of nutrition onto health. Now when we intend on answering **Q1** it seems reasonable to start with the queried variable first, mobility in this case. Since we know that M is an effect of H with $\gamma > 0$ we expect the below average mobility to be explained by an already below average health value. Indeed, this expectation is met since $H_H < \mu_H$. Traversing the chain further to the causes of H, which are A, N, we observe two different scenarios. Since A is above average as Hans is an elderly person ($H_A > \mu_A$) and $\beta < 0$ we can conclude that H_A is definitely an explanation for H_H whereas N with $\gamma > 0$ is actually a countering factor since Hans has a good diet ($H_N > \mu_N$) beneficial to his health. In summary, by exploiting the knowledge on $\mathcal{M}$ we have arrived at a causal explanation that can be pronounced in natural language as:

Explanation 1 (for Q1) *"Hans's Mobility is bad because of his bad Health which is mostly due to his high Age although his Food Habits are good."*

Explanation 2 is indeed an explanation as required by the definition of (Dennett, 1989) since it is an answer to the why-question concerning Hans' mobilitity. Furthermore, it is a causal explanation since the used coefficients for deriving the explanation are based on SCM $\mathcal{M}$ that satisfies the assumptions from Sec.2 thus qualifying the coefficients as causal effects. Our above explanation captures two prominent modes of human reasoning, namely both the existence and the "strength" of a causal relation. In the following we will capture and formalize our intuition that allowed us to derive Exp.2, which in turn allows us to compute such causal explanations automatically.

When reflecting on the actual knowledge used in our argument above, then we realize that we can abstract away four key aspects: (i) that there is a relative notion in the why-question Q_M like "why ... bad?" that implicitly compares an individual (here, Hans) to the remaining population of patients, (ii) the causal graph provides the structure of the explanation by following any previously unexplained directed path to the target effect (here, mobility), (iii) the causal effect for any pair allows us to assert whether the observed values for that pair are "surprising" in that they are consistent with the mechanisms of the data-generating process or not, and (iv) that some causal effects are more important or influental than others (here, age versus nutrition w.r.t. health). Following this reflection we define sth. called *causal scenario* that will cover (i-iii) as point (iv) will be covered separately.

Definition 5 *As before let $V_i, V_j \in \mathbf{V}$ and α denote the ACE from V_j onto V_i and μ_i, μ_j are the averages of our data sample. The tuple $\mathbf{C}_{i,j} := (\alpha, v_i, v_j, \mu_i, \mu_j)$ is called a causal scenario.*

With this convenient notation at hand, we are ready to abstract the logic of our explanation to general rules. For this we will make use of first-order logic.

Definition 6 (Explanation Rules) *Let $\mathbf{C}_{i,j}$ denote a causal scenario, $R_i \in \{<, >\}$ be a binary ordering relation and $\boldsymbol{\alpha}_i^{\mathrm{pa}}$ be the set of all absolute parental ACEs onto V_i. We define FOL-based rule functions as followed by indicating for each rule ERx: $(\mathbf{C}_{i,j}, R_1, R_2, \boldsymbol{\alpha}_i^{\mathrm{pa}}) \mapsto \{-1, 0, 1\}$ how the causal relation $V_i \leftarrow V_j$ satisfies that rule.*

(ER1) If $R_1 \neq R_2$, then: $((\alpha < 0) \wedge \delta_1) \vee ((\alpha > 0) \wedge \delta_2)$

(ER2) If $R_1 \neq R_2$, then: $((\alpha > 0) \wedge \delta_1) \vee ((\alpha < 0) \wedge \delta_2)$

$$\textit{with } \delta_1 := R_2(v_j, \mu_j) \wedge R_1(v_i, \mu_i) \textit{ and } \delta_2 := R_2(v_j, \mu_j) \wedge R_2(v_i, \mu_i)$$

And as an extra, "modifier" rule in the case where $|\boldsymbol{\alpha}_i^{\text{pa}}| > 1$ we simply consider the parent with the highest ACE absolutely, $V_k^* = \arg\max_{V_k \in \text{pa}(V_i)} \boldsymbol{\alpha}_i^k$, as the most dominant cause. Concretely, ER3 does not test a new logical condition but modifies the pronunciation of whichever of ER1/2 fired for dominant parent V_k^* by prefixing "mostly" (see Tab. 1). For all other parents, ER3 outputs 0.

Since our rules only need qualitative knowledge on the causal effects (i.e., we simply test the sign of the coefficients) it is possible to use techniques from *partial identification* which allows for bounding causal effects using fewer necessary assumptions at the price of exact estimation (Balke and Pearl, 1994). These two plus one rules build the foundation for our new type of explanation. Having the actual relation R as a return argument of each of the rules allows for a fine-grained explanation. In a nutshell, it allows to extend a statement "V_i because of V_j" to a more detailed one like "V_i because of V_j *being low*". The general pronunciation scheme for the the three rules which we name excitation (*ER*1), inhibition (*ER*2), and preference (*ER*3) are summarized in Tab.1. The

*ER*1	Excitation	"V_i because of V_j [being low/high]"
*ER*2	Inhibition	"V_i although V_j [is low/high]"
*ER*3	Preference	"mostly" + *ER*1 or *ER*2 pronunciation

Table 1: **Pronunciation Scheme.** Right shows the natural language reading of a rule's activation.

pronunciation of the details to the relation e.g. "low"/"high" is context-dependent in that these words might need to replaced with adequate/corresponding words suitable for the context. To elaborate, "the mountain top is cold because of the high altitude" is fine, while "the remaining car fuel is low because of the driver's bad driving style" requires the context-adaptation (what was "low" previously is "bad" in this case). Another noteworthy detail to the CAT properties is the property of *non-repeating causes within explanations* which reduces redundancy. Consider for instance our lead example on Hans's mobility (Exp.2), the SCM suggests that N can also be explained by A, since $A \rightarrow N$. However, the corresponding CAT output does not give this reason because of the aforementioned property which ensures that redundancy is being avoided. I.e., in the explanation step before we actually explain H using both A and N, since $\{A, N\} \rightarrow H$, therefore, making it irrelevant for the question to explain the relation between the parents (A, N). While we provided intuition on the derivation of these basic FOL rules alongside the "Causal Hans" example, we now additionally motivate the namings "excitation", "inhibition" and "preference". We took inspiration from *neuroscience*, where the former two terms relate to the way neurons interface with each other using their synaptic-dendric connections (He and Cline, 2019). The last term is a term to propose "relativity" and thus a preference for one cause over the other. Returning to our derivation, to now show how these rules can generate something like Exp.2, we will present our actual algorithm definition. We define the CAT-algorithm as:

Definition 7 (CAT) *Like before let $Q_i, \mathcal{M}$ be a valid why-question and some SCM estimate respectively. Further, let $\mathbf{D} \in \mathbb{R}^{n \times |\mathbf{V}|}$ denote our n-samples data set. We define a recursion*

$$\mathbf{E}(Q_i, \mathcal{M}, \mathbf{D}) = (\bigoplus_{V_k \in \text{pa}(V_i)} ER(V_i \leftarrow V_k), \bigoplus_{V_k \in \text{pa}(V_i)} \mathbf{E}(Q_k, \mathcal{M}, \mathbf{D})) \tag{1}$$

where $ER(V_i \leftarrow V_k)$ *is a shorthand for* $\{ERx(\mathbf{C}_{i,j}, R_1, R_2, \boldsymbol{\alpha}_i^{\text{pa}})\}_{i=1,2,3}$ *that are given through* $(\mathcal{M}, \mathbf{D})$ *and* $\bigoplus_{i=1}^{n} v_i = (v_1, \ldots, v_n)$ *denotes concatenation. The recursion's base case is being evaluated at the roots of the causal path to* V_i*, that is, for some* $V_k \in \mathbf{V}$ *with a path* $V_k \to \cdots \to V_i$ *we have*

$$\mathbf{E}(Q_k, \mathcal{M}, \mathbf{D}) = \emptyset. \tag{2}$$

We call $\mathbf{E}(Q_i, \mathcal{M}, \mathbf{D})$ *the CAT output for* V_i *based on* $(\mathcal{M}, \mathbf{D})$.

How to read the articulation rules and recursion. The rules ER1–ER3 in Def. 6 are best read as graph-traversal checks. At each directed edge $V_j \to V_i$, one evaluates the sign of α and whether v_j and v_i are above or below their respective means. ER1 fires when the observed state of V_j is *consistent* with the sign of α causing the observed state of V_i (excitation). ER2 fires in the complementary case, i.e., the effect of V_j on V_i is working *against* the observed direction (inhibition). ER3 then selects the strongest excitatory or inhibitory parent. The recursion in Def. 7 proceeds like a depth-first search over the DAG: starting from the queried variable V_i, it collects the ER outputs for all parent edges $(V_k \to V_i)$ and then recurses into each parent V_k in turn, stopping at roots ($\text{pa}(V_k) = \emptyset$). The result is a nested tuple read off as a natural-language sentence following Tab. 1.

CAT: Procedural Summary (**Input:** why-question Q_i, SCM $\mathcal{M}$, data $\mathbf{D}$; **Output:** $\mathbf{E}(Q_i, \mathcal{M}, \mathbf{D})$)

1. **Base case:** if $\text{pa}(V_i) = \emptyset$, return $\emptyset$.
2. **For each** $V_k \in \text{pa}(V_i)$: compute scenario $\mathbf{C}_{i,k}$ from $(\mathcal{M}, \mathbf{D})$, then evaluate ER$1, 2, 3$ on $\mathbf{C}_{i,k}$.
3. **Recurse:** call $\mathbf{E}(Q_k, \mathcal{M}, \mathbf{D})$ for each V_k.
4. **Return** the concatenation of all ER outputs and recursive results.

For humorous simplicity we refer to the general family and any particular set of CAT outputs as CATs. The above algorithm is a simple recursion that traverses all possible directed, causal paths to the target variable checking each of the rules *ERx* thus constructing a unique code that maps to a unique answer following our previous pronunciation scheme. Since CAT answers a why-question, which is counterfactual by nature, and does so by using qualitative knowledge of the SCM, which encodes counterfactual knowledge, we can generally classify CAT as a *counterfactual*-type of explanation in the broader scope of conceptually distinct ideas in causal explainable AI. In the following we show how to reconstruct the initial Causal Hans Example using Def.7. To return one last time to our running example, we apply the recursion step-by-step now. For Q_M (corresponding to **Q1**) we get (note: best viewed in color, step-wise color-coded presentation):

$$\begin{aligned}
&\mathbf{E}(Q_M, \mathcal{M}, \mathbf{D}) \\
&= ((ER1 = -1), \textstyle\bigoplus_{V_k \in \{A,F\}} \mathbf{E}(Q_H, \mathcal{M}, \mathbf{D})) \\
&= (\ldots, (((ER1 = 1, ER3 = 1), \mathbf{E}(Q_A, \mathcal{M}, \mathbf{D})), ((ER2 = 1), \mathbf{E}(Q_F, \mathcal{M}, \mathbf{D})))), \\
&= (\ldots, ((\ldots, \emptyset), (\ldots, \emptyset))).
\end{aligned}$$

So the recursion result is $H \to M : (ER1 = -1, ER2 = 0, ER3 = 0), A \to H : (ER1 = 1, ER2 = 0, ER3 = 1), F \to H : (ER1 = 0, ER2 = 1, ER3 = 0)$. This result *uniquely* identifies the human-readable pronunciation of our causal explanation in Exp.2. For the graphical interpretation refer to Fig.4 which highlights the recursive traversal through each pair of parents while avoiding redundancy through duplicate paths.

4. Augmenting CAT to Account for Time

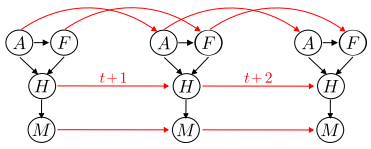

Figure 1: Truncated full-time graph (Hans Ex.). Red: delayed effect. (Best viewed in color.)

The temporal extension of CAT presented in this section is a *structural* extension of the articulation principle to time-series and agentic settings. It demonstrates that the core ideas of CAT—qualitative causal reasoning grounded in sign and magnitude of effects—transfer naturally to dynamic contexts. We do *not* claim full practical optimization for arbitrarily long temporal horizons; scalability and summarization over long horizons are explicitly identified as directions for future work. We expand the Causal Hans example by incorporating a time dimension into the analysis. As shown in Fig. 1, our new causal time graph includes both immediate effects (black) and delayed self-effects (red). Our synthetic dataset, consisting of 10,000 records with 50 time steps each, shares a similar structure with the original dataset. Initially, the age distribution P_A is uniformly distributed between 30 and 80, with each time step increasing by 1. P_F is defined as $P_F = 0.5 \cdot P_A$, P_H is computed using $P_H = -0.2 \cdot P_A + 0.6 \cdot P_F$, and P_M is established as $P_M = 0.5 \cdot P_H$. At each time step, values are drawn from the distribution, multiplied by 0.4, and combined with the previous time step's influence, calculated with 0.6. To maintain consistency, each distribution's mean is determined, and noise is introduced through an added $\mathcal{N}(0, \mu^P \cdot 0.03)$.

Generalization 1 (Why Question) *For individual i and their instance $x_t^i \in Val(X_t)$ of any variable $X_t \in \mathbf{V}$ in SCM $\mathcal{M}$ at time point t, let ϕ^{X_t} be a population statistic (e.g., mean, 10% percentile). A why-question $Q_{X_t^i} \stackrel{\text{def}}{=} R(x_t^i, \phi^{X_t})$, with $R \in \{<, >\}$, is true if $Q_{X_t^i}$ holds.*

Generalization 2 (Causal Scenario (CS)) $C_{XY} \stackrel{\text{def}}{=} (\alpha_{X \to Y}, x_t, y_k, \phi^{X_t}, \phi^{Y_k})$ *is called a CS.*

Generalization 3 (Explanation Rules) *Let C_{XY} denote a causal scenario. Given a sign function $s(x) \in \{-1, 1\}$, a binary ordering relation $R_i \in \{<, >\}$, we define FOL-based rule functions $ERi(\cdot) \in \{-1, 0, 1\}$, indicating how the causal relation $X \to Y$ satisfies each rule. $ERi(\cdot)$ evaluates both Fundamental Rules and situational Complementary Rules, which can be added. Here, t and k are time steps, where $t \geq k$*

(ER1) If $R_1 \neq R_2$, then: $((s(\alpha_{X \to Y}) < 0) \wedge \delta_1) \vee ((s(\alpha_{X \to Y}) > 0) \wedge \delta_2)$

(ER2) If $R_1 \neq R_2$, then: $((s(\alpha_{X \to Y}) > 0) \wedge \delta_1) \vee ((s(\alpha_{X \to Y}) < 0) \wedge \delta_2)$

$$\text{with } \delta_1 \stackrel{\text{def}}{=} R_2(x, \phi^{X_t}) \wedge R_1(y, \phi^{Y_k})) \text{ and } \delta_2 \stackrel{\text{def}}{=} R_2(x, \phi^{X_t}) \wedge R_2(y, \phi^{Y_k}))$$

Definition 8 (Explanation Tree) *An Explanation Tree is a directed acyclic graph (DAG) $\mathcal{T}$. The root node $r \in \mathcal{N}$ represents the variable of interest (valid Why-Question). Each edge $(u, v) \in \mathcal{E}$ is directed from parent node $u \in \mathcal{N}$ to child node $v \in \mathcal{N}$, such that in the retrospective case, the children of a node in the tree represent the causally explanatory variables of their parent node. In the anticipative case, the relationship is reversed, with the child nodes being explained by their parent node in the tree.*

As before, we focus on the patient Hans. His data can now be represented at each time step t as a tuple $X_t = (X_t^A, X_t^N, X_t^H, X_t^M)$. Valid questions are defined according to Gen. 1, and the causal scenario (Gen. 2) aligns with the original concept. The Explanation Rules (Gen. 3) have been adapted to accommodate time series data while still preserving the underlying idea that these rules encode causal relationships. Notably, variables X and Y can now originate from two distinct time steps. While the original algorithm used a linked list as its fundamental data structure and produced nested sentences for explanations, we have chosen to extend it with a tree data structure (Def. 8). As the number of explanatory variables increases over time and for larger SCMs, nested explanations inevitably lead to confusion. Moreover, our switch to the new data structure implies that we aim to provide one-sentence explanations for each variable to be explained. As we will see later, this also has advantages for summarizing causal relationships over time or identifying changing relationships. To construct the Explanation Tree, the CAT recursion must be adjusted accordingly (see Def. 9).

Definition 9 (CAT (+Time)) *Let $\mathrm{Q_R}$ be the root node of a valid why-question and $\mathfrak{M}$ a set of proxy SCMs for distinct contexts. Moreover, let $D \in \mathbb{R}^{n\times|\mathbf{V}|}$ represent our dataset and $K \in \mathbb{N}$ denote the maximum recursion depth, with the starting depth being $j = 0$. In the first iteration, $\mathrm{Q_X} = \mathrm{Q_R}$. We define a recursion as follows:*

$$\mathbf{E}(\mathrm{Q_R}, \mathrm{Q_X}, \mathfrak{M}, D, j) = \Big(\bigoplus_{Z\in\Phi_1} EI(Z, X_t), \bigdiamond_{Z\in\Phi_1\setminus\Phi_2} \begin{cases} \mathbf{E}(\mathrm{Q_X}, Z, \mathfrak{M}, D, j+1), & \text{if } j < K \\ \emptyset, & \text{else} \end{cases} \Big) \quad (3)$$

$\bigoplus$ attaches new nodes to the given node depending on the used case. Explanation Indicators, $EI(\cdot)$, resolves all ERi and stores the time step t, variable names and context of the nodes. The iterator $\bigdiamond$ maintains the recursion over included nodes. The Φ_1 further select the appropriate approx. SCM from $\mathfrak{M}$ based on the dataset, time step, and variable (context's). The set Φ_2 prevents from reattaching duplications by checking for existing time t and variable name combinations.

Definition 10 (Retrospective) *If $\Phi_1 \overset{\text{def}}{=} (\mathrm{Pa}_t^X \cup \mathrm{Pa}_{t,\theta}^X \cup \mathrm{Pa}_{t,n}^X)$ and $\Phi_2 \overset{\text{def}}{=} \mathrm{De}^{\mathrm{Q_R}}$ where $\mathrm{Pa}_{t,\theta}^X \overset{\text{def}}{=} \{Z \in \mathrm{Pa}_t^X \mid |\alpha_{Z\to X}| > \theta\}$ and $\mathrm{Pa}_{t,n}^X \overset{\text{def}}{=} \{\mathrm{top}n(Z \in \mathrm{Pa}_t^X, |\alpha_{Z\to X}|)$ then we call $\mathbf{E}$ retrospective. Here $\theta \geq 0$ is an effect-magnitude threshold filtering out negligible parents, and $n \in \mathbb{N}$ bounds the number of retained parents; both are user-chosen scope-control parameters.*

Definition 11 (Anticipative) *If $\Phi_1 \overset{\text{def}}{=} (\mathrm{Ch}_t^X \cup \mathrm{Ch}_{t,\theta}^X \cup \mathrm{Ch}_{t,n}^X)$ and $\Phi_2 \overset{\text{def}}{=} \mathrm{An}^{\mathrm{Q_R}}$ where $\mathrm{Ch}_{t,\theta}^X \overset{\text{def}}{=} \{Z \in \mathrm{Ch}_t^X \mid |\alpha_{X\to Z}| > \theta\}$ and $\mathrm{Ch}_{t,n}^X \overset{\text{def}}{=} \{\mathrm{top}n(Z \in \mathrm{Ch}_t^X, |\alpha_{X\to Z}|)$ then we call $\mathbf{E}$ anticipative. The parameters θ and n play the same scope-control role as in Def. 10.*

Def. 9 is used to construct an Explanation Tree $\mathcal{T}$ for a valid question. Additionally, a distinction is made between the retrospective and anticipatory cases. Given time series and a causal graph, we can provide explanations for the emergence of our variable of interest at a specific point (retrospective reasons, see Def. 10). Additionally we can include causal effects as explanatory factors, e.g., when we aim to explain behavior (anticipative effects, see Def. 11). Furthermore, these case differentiations are designed to narrow down the possible number of explanatory variables and select the appropriate SCM among different options for a given point in time. For our time series Causal Hans example, we currently focus on one context (SCM) and retrospective explanation. The data structure now allows for further indicators or manipulations to be performed directly on the tree.

A sequence indicator, which helps to mark sequences of consistent causal relationships, is of crucial importance for summarization and identification of changing relationships. Masking variables (e.g., Age, as the causal explanation is generally given) or trimming down the tree to individual path explanations are also possible. To implement the retrospective sequence indicator, we need to define a function $f(\mathcal{T})$ that iterates separately over all endogenous variables of the SCMs in $\mathcal{T}$ in descending temporal order. If direct child nodes have the same Explanation Rule indicators and remain stable over time, a unique ID is assigned to the sequence. Leaves should be excluded, since they are just explanatory and not to be explained. In the following, we provide an example illustrating human-readable causal explanations in response to the question "Why is Hans' Health below average?" with the recursion depth set to 2. The varying explanations for Hans' nutrition and the perfectly summarizable explanation for his health are highlighted in color. It should be noted once more that effects from the past are always positive, and as in the original causal Hans graph, the effect of age on nutrition is also positive, whereas age has a negative impact on health. The corresponding Explanation Tree is shown in Fig. 2. See the appendix for a more detailed example.

Explanation 2 (CAT output for Q1 (+Time)) *"Hans' health has consistently been below average over the last two years, mostly because his low Health persistently one year prior and because of his high Age in the referenced year, despite his high Nutrition in the referenced year. His diet was above average this year, due to his high Nutrition one year before and his high Age in the same year. Last year, his diet was above average, because of his high Age in the same year, despite his low Nutrition one year before."*

5. Related Work and Conclusions: Contextualizing CAT with Existing Causal XAI

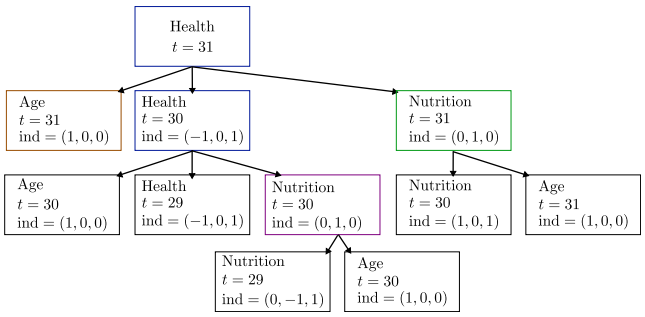

Figure 2: Sub-tree with colored sequences, question: "Why is Hans' Health below average?". (Best viewed in color.)

A great body of work within deep learning has provided visual means for explanations of how a neural model came up with its decision i.e., importance estimates for a model's prediction are being mapped back to the original input space e.g. raw pixels in the arguably standard use-case of computer vision (Sundararajan et al.; Selvaraju et al., 2017; Schulz et al., 2020). To circumvent explanations that are like "children that are only able to point fingers", Stammer et al. (2021) proposed a neuro-symbolic explanation scheme to revise behavior from learned models in an interactive loop following the framework of (Teso and Kersting, 2019). On the causal end, (Schwab and Karlen, 2019) proposed a model-agnostic approach that can generate explanations following the idea of Granger causality (which is very different from Pearlian causality as it captures "temporal relatedness" which holds in their setting as input precedes output). On the Pearlian side of explanations, the computation of Causal Shapely Values (Heskes et al., 2020) or the LEWIS framework (Galhotra et al., 2021) are explainers for numerical attribution that capture important distinctions within

causality such as direct vs. indirect causes or the necessity-sufficiency distinction of causes. Closest to our work on a semantic level within Pearl's causal framework are arguably works on fairness (Kusner et al., 2017; Plecko and Bareinboim, 2022). For instance, Karimi et al. (2020) investigated how to best find a counterfactual that flips a decision of interest e.g. an applicant for a credit is rejected and the question is now which counterfactual setting (changes to the applicant) would have resulted in a credit approval. Considering unit-level instead of population-level causality, our work can compare to the definitions of Halpern (2016) for "actual causation", where the key difference lies in the *relativity* of our explanation approach to a given sample population in addition to the overall less philosophical approach to causal explanations that shows in both how we generate the explanations and then use them for learning. To conclude the main part of this paper with more than pure theory, we aim to compare CAT (+Time) with other state-of-the-art (SotA) causal explanation methods and, of course, our original introduction of the fundamentals of CAT vanilla. Since as of writing, no method is available that generates both SCM-based explanations over time, we must limit ourselves to a very asymmetric comparison. Furthermore, each method is based on different assumptions and is designed for different purposes. Our method relies on explaining through the causal effect *direction* and *strength*, rather than the combination of feature value and structural influence of individual attributes. For evaluation purposes, we are interested in our running patient example "Why is Hans' Mobility below average?". For the comparison, we have adapted the synthetic Causal Hans time series dataset so that the counterfactual-type explanations Galhotra et al. (2021) and Causal Shapley Values Heskes et al. (2020) can be used similarly. The new dataset is designed as a classification problem and contains the categorization 'below average' (0) or 'above average' (1) for all variables across all patients and time steps. In line with our causal graph 1, we have added delayed variables and immediate variables as 'lagged' (prefix 'Lag_') and 'non-lagged' respectively in each sample. The counterfactual explanation method LEWIS enables the calculation of sufficiency- (Suf), necessity- (Ne), and NeSuf-Score(s) to provide explanations at various levels (local, contextual, global) for a binary classification (e.g., credit approval). Depending on the feedback (0 or 1), a different value (Suf or Ne) is important for reversing the feedback with a certain probability (calculated score). In terms of our classification problem, this signifies that a prediction of 0 (below average) is depicted by the Sufficiency Score of each variable, illustrating how each could potentially contribute to a shift in mobility. As we are interested in the extent to which the current value is used for explanation, we choose the alternative category as the starting point and calculate a transition to the current category. Contrary to our method, we obtained diverse explanatory variables (and values) for the local explanations here. Fig. 3 (left side) presents the Suf Scores for 100 patients with prediction below average mobility. For our dataset, Suf Scores in the range of 0 were often calculated across all variables for several patients. In isolated cases, variables received differing Suf Scores, with the most variation in 'Lag_Mobility'. Ultimately, the method is not designed to distinguish direct from indirect effects or to generate consistent explanation variables across different patients. Moreover, the evaluation of whether a category of a feature has a positive or negative effect is only possible through trial and error, insofar as there is no prior knowledge about the categories. To compare with the Causal Shapley Values, we used a similar setup. The Causal Shapley Values differ significantly from other Shapley Values (Aas et al., 2020; Janzing et al., 2019; Frye et al., 2021), as they are able to quantify the influence of direct and indirect variables differently. The inclusion of a partial order of the real causal structure underlying the data limits the permutation possibilities for calculation. The authors have introduced both symmetric and asymmetric Causal Shapley Values. Since asymmetric values tend to place more

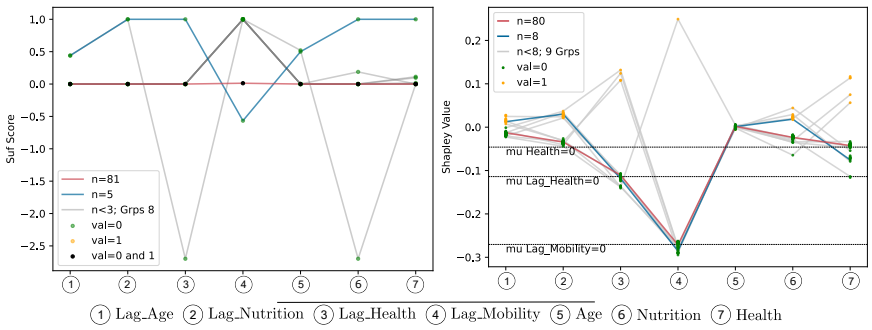

Figure 3: Suf. score and Shapley vals. for 100 random patients with below avg. mobility predicted. Same sequences are grouped together (group size n). (Best viewed in color.)

weight on the root element ('Lag_Age' in our case) in chain graphs, and we are more interested in direct and indirect effects, we used symmetric values for our experiment, which include both influences in the calculations and distribute the effect. For testing purposes, we provided a partial order corresponding to the temporal sequence of our causal graph ([{'Lag_Age', 'Lag_Nutrition', 'Lag_Health', Lag_Mobility'},{Age, Nutrition},{Health}]) and plotted the Causal Shapley Values for 100 patients with below average mobility in Fig. 3. It is evident that 'Lag_Mobility' exerts a significant influence. The direct effect of 'Health' is likely diminished in this partial order due to its involvement in many indirect effects. Notably, within this order, 'Lag_Health' assumes greater importance than 'Health' in the predictions. Whether a (truly direct) variable generally has a positive or negative effect on the target variable in this case can be straightforwardly determined by comparing it with the complementary group 'above average.' For our case, this roughly corresponds to a reflection along the x-axis and an inversion of colors. The resulting values can be used to explain individual predictions, but it requires considerable additional effort in the analysis to focus exclusively on direct effects. Causal Shapley Values do incorporate information on the causal structure but neither in conjunction with mechanism information nor in the same manner for the presented direct articulation through CAT.

Conclusive thoughts and future work. We've presented a conceptually original approach to causal explanations based on SCMs. Throughout this work, *faithfulness* is understood in the SCM sense: an explanation is faithful if it is derived directly from the structural parameters and graphical mechanisms of the SCM, rather than learned from a mapping from inputs to rationales. This way, our 'SCM-faithfulness' corresponds in intuition to the usual encounter of this word within causal discovery literature, where a distribution is faithful to a graph, whenever the distribution does not introduce independences that the graph's structure is not aware of. The empirical components—human alignment study, causal discovery regularization, and temporal examples—are designed to validate that CAT behaves coherently in practice under this notion of faithfulness. Naturally, there is a lot of opportunity for future work. We presented our theory with an immediate extension to time, however, recursive explosion w.r.t. number of time steps measured, that is, how to handle redundancy and the scope of an explanation is ongoing issue. On a conceptual note, making quantitative use of the knowledge on causal effects instead of purely qualitative knowledge would likely allow for more expressive explanations. Finally, we believe that interactive approaches to explanations are a promising paradigm for CAT.

Acknowledgments

We thank the reviewers for their constructive and thoughtful feedback, which has helped improve the presentation of this work. We thank our colleagues at TU Darmstadt and TU Eindhoven including our causality team: Moritz Willig, Florian Peter Busch, Tim Woydt, Jonas Seng, Nicholas Tagliapietra. The TU Eindhoven authors received support from their Department of Mathematics and Computer Science and the Eindhoven Artificial Intelligence Systems Institute. This work has benefited from the early stages of the fundings by the German Research Foundation (DFG) under Germany's Excellence Strategy— "Reasonable AI" (EXC-3057) and "The Adaptive Mind" (EXC-3066). For the user study evaluation we thank Constantin Rothkopf. As a final acknowledgment, Matej Zečević dedicates this work in loving memory to Nika Ovcharova and Kuki Zečević.

References

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values, 2020.

Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty Proceedings 1994*, pages 46–54. Elsevier, 1994.

Stephan Bongers, Patrick Forré, Jonas Peters, and Joris M Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915, 2021.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. *arXiv preprint arXiv:2007.01754*, 2020.

Daphna Buchsbaum, Sophie Bridgers, Deena Skolnick Weisberg, and Alison Gopnik. The power of possibility: Causal learning, counterfactual reasoning, and pretend play. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2012.

Ruth MJ Byrne. Counterfactual thought. *Annual review of psychology*, 67:135–157, 2016.

Tathagata Chakraborti, Sarath Sreedharan, Yu Zhang, and Subbarao Kambhampati. Plan explanations as model reconciliation: Moving beyond explanation as soliloquy. *IJCAI*, 2017.

Clive R Charig, David R Webb, Stephen Richard Payne, and John E Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *Br Med J (Clin Res Ed)*, 1986.

Max Chickering, David Heckerman, and Chris Meek. Large-sample learning of bayesian networks is np-hard. *JMLR*, 2004.

Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1282–1289. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/cobbe19a.html.

Kahneman Daniel. Thinking, fast and slow, 2017.

Daniel C Dennett. *The intentional stance*. MIT press, 1989.

Christopher Frye, Colin Rowat, and Ilya Feige. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability, 2021.

Sainyam Galhotra, Romila Pradhan, and Babak Salimi. Explaining black-box algorithms using probabilistic contrastive counterfactuals. In *Proceedings of the 2021 International Conference on Management of Data*, pages 577–590, 2021.

Tobias Gerstenberg, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. How, whether, why: Causal judgments as counterfactual contrasts. In *CogSci*, 2015.

Tobias Gerstenberg, Matthew F Peterson, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. Eye-tracking causality. *Psychological science*, 28(12):1731–1744, 2017.

Alison Gopnik. Scientific thinking in young children: Theoretical advances, empirical research, and policy implications. *Science*, 2012.

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and interpretable models in computer vision and machine learning*, pages 39–80. Springer, 2018.

Thomas L Griffiths and Joshua B Tenenbaum. Optimal predictions in everyday cognition. *Psychological science*, 17(9):767–773, 2006.

Joseph Y Halpern. *Actual causality*. MiT Press, 2016.

Masasi Hattori. Probabilistic representation in syllogistic reasoning: A theory to integrate mental models and heuristics. *Cognition*, 157:296–320, 2016.

Hai-yan He and Hollis T Cline. What is excitation/inhibition and how is it regulated? a case of the elephant and the wisemen. *Journal of experimental neuroscience*, 13:1179069519859371, 2019.

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

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11 (5), 2010.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem, 2019.

John R Josephson and Susan G Josephson. *Abductive inference: Computation, philosophy, technology*. Cambridge University Press, 1996.

Amir-Hossein Karimi, Julius Von Kügelgen, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in Neural Information Processing Systems*, 33:265–277, 2020.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

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

David A Lagnado, Tobias Gerstenberg, and Ro'i Zultan. Causal responsibility and counterfactuals. *Cognitive science*, 37(6):1036–1073, 2013.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 2019.

John McCarthy. What is artificial intelligence? 1998.

John McCarthy and Patrick J Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pages 431–450. Elsevier, 1981.

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

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Nancy J Nersessian. In the theoretician's laboratory: Thought experimenting as mental modeling. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 1992.

Judea Pearl. *Causality*. Cambridge university press, 2009.

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

Derek C Penn and Daniel J Povinelli. Causal cognition in human and nonhuman animals: A comparative, critical review. *Annu. Rev. Psychol.*, 58:97–118, 2007.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference*. The MIT Press, 2017.

Oskar Pfungst. *Clever Hans:(the horse of Mr. Von Osten.) a contribution to experimental animal and human psychology*. Holt, Rinehart and Winston, 1911.

Drago Plecko and Elias Bareinboim. Causal fairness analysis. *arXiv preprint arXiv:2207.11385*, 2022.

Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 765–804. 2022.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. Restricting the flow: Information bottlenecks for attribution. *arXiv preprint arXiv:2001.00396*, 2020.

Patrick Schwab and Walter Karlen. Cxplain: Causal explanations for model interpretation under uncertainty. *arXiv preprint arXiv:1910.12336*, 2019.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

Herbert A Simon. Modeling human mental processes. In *Western joint IRE-AIEE-ACM computer conference*, 1961.

Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951.

Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *CVPR*, 2021.

Christian Steinruecken, Emma Smith, David Janz, James Lloyd, and Zoubin Ghahramani. The automatic statistician. In *Automated Machine Learning*, pages 161–173. Springer, Cham, 2019.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*. PMLR.

Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *AIES*, 2019.

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *NeurIPS*, 2018.

Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric Xing. Learning sparse nonparametric dags. In *International Conference on Artificial Intelligence and Statistics*, pages 3414–3425. PMLR, 2020.

Appendix A. Empirics: Vanilla CATs

Does CAT Output Sensible Explanations? To get an understanding of whether CAT outputs are sensible beyond the running example we have shown, we have conducted a user study with $N = 22$ human subjects that had to judge the qualitative causal structure of four "daily-life" examples using a questionnaire specifically designed to provide us with the data necessary for constructing causal graphs representative of what the participants think about the presented concepts. The first question to answer is: how did we construct the graph estimates from human data? In Fig.5 we show two ways that we considered: the "Mode" refers to the scheme where we simply look at the different

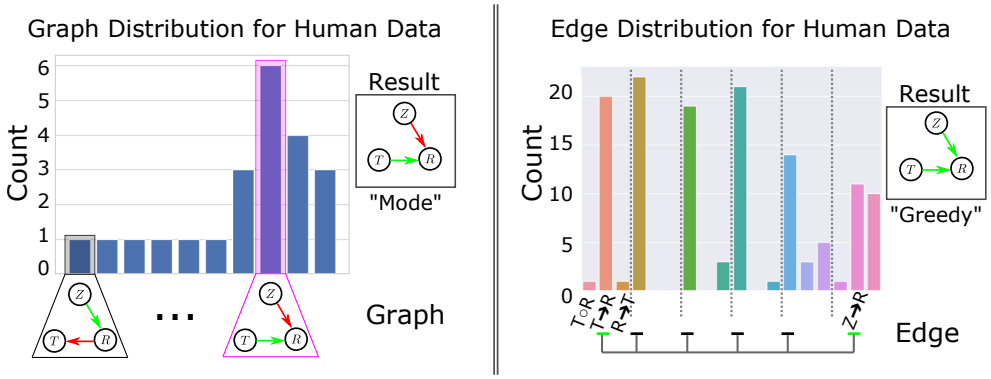

Figure 5: **Measuring Agreement Between Beliefs of Human Subjects Regarding Causal Graphs.** Left, the graph estimate is the mode of the distribution of all predicted causal graphs. Right, greedily pick each edge of the graph. (Best viewed in color.)

graphs and take the *most frequently reoccurring* graph as representative of the population, or the "Greedy" approach where we look at the frequency at which edges are predicted and then simply construct a graph from greedily taking the *most probable edge each time*. Greedy comes at the cost that the predicted graph is not necessarily within the populaton. With the human causal graphs at hand, we now investigate our intial question about CAT. For brevity, we will only highlight the most important key observations with a prolonged discussion being provided in the Appendix: Observation **(i)** the CATs that we generate from the acquired causal graphs are sensible in the sense that they lie close (or are even identical) to the apriori expectation of the study (the proposed ground truth). Observation **(ii)** we observe a systematic approach and thereby non-random approach to edge-/structure-selection by the subjects. Furthermore, there are only a few clusters even with increasing hypothesis space. Both the systematic manner and the tendency to common ground are evidence in support prior evidence that SCMs are a suitable representation for human causal modelling. Observation **(iii)** we observe that the increase in hypothesis/search space (i.e., more variables) comes with an increase in variance. This variance increase can be argued to be due to the progressive difficulty of inference problems as well as decreased levels of attention and potential fatigue across the duration of the experiment, and observation **(iv)** some subjects implicitly assume a notion of time by assuming a cyclic relationship between e.g. treatment and recovery, where the subject likely thought in terms of 'increasing treatment increases the speed of recovery *which subsequently* feeds back into a decrease of treatment'.

What Can CAT Reveal About Graph Learning Methods? Having established the CAT algorithm as a sensible way for producing explanations, the natural next step is to conisder how we can incorporate CAT into learning. To this end, we start by considering CAT outputs generated from graphs learned by popular graph learning methods. Induction of inter-variable relationships based on available data, especially of directed acyclic graphs (DAGs), is paramount in causality (Pearl, 2009). Unfortunately, due to the combinatoric nature of the problem setting, learning DAGs from data is recognized to be an NP-hard problem (Chickering et al., 2004). However, several works have tackled this difficult problem and one solution for learning linear DAGs came from a method called NOTEARS (abbreviated NT, Zheng et al. (2018)) who were able

to re-formulate the traditional view into a continuous shape such that any non-convex optimization can be applied for the graph estimation problem. The authors propose the general formulation, $\min_{\mathbf{W}\in\mathbb{R}^{d\times d}} \quad f(\mathbf{W}) \quad \text{subject to} \quad h(\mathbf{W}) = 0$, where f is a data-based score, e.g. a regularized least-squares loss is applied assuming a sparse linear model (possibly SCM). That is $f(\mathbf{W}) = ||\mathbf{X} - \mathbf{X}\mathbf{W}||_F^2 + ||\mathbf{W}||_1$, and h is a smooth function with a kernel (or null space) that only contains acyclic graphs, $h(\mathbf{W}) = 0 \iff \mathbf{W}$ is acyclic. Different variations of the same continuous counting mechanism using this acyclicity constraint have been proposed, e.g., Zheng et al. (2020) proposed $h(\mathbf{W}) = \text{tr}(e^{\mathbf{W}\circ\mathbf{W}}) - d$ while Yu et al. (2019) proposed $h(\mathbf{W}) = \text{tr}[(\mathbf{I} + \mathbf{W}\circ\mathbf{W})^m] - m$. Unfortunately, both suffer from cubic runtime-scalability in the number of graph nodes, $O(d^3)$. While the aforementioned works have focussed on data originating from (non-linear transformation) of linear SCM, there exists yet another sub-class of DAG-learning methodologies that focuses on more general causal inference. Ke et al. (2019) made use of interventional data to update their graph estimate while using masked neural networks to mimic the structural equations. Brouillard et al. (2020) follows the same idea of leveraging causal information, e.g. interventional data, for overcoming identifiability issues while staying close to the continuous optimization formalism introduced by NT. Returning to our question: we looked at different data sets including different graph learners and for each combination generated their respective CAT output. We considered several different why-questions for each of the four data sets: data set for the Causal Hans example, weather forecast (W, real world, Mooij et al. (2016)), mileage (M, synthetic), and recovery (R, real world, Charig et al. (1986)). To avoid cluttering in the main text we have moved the relevant tables and figures to the the Appendix where we also provide an extended account, here we highlight the most important insights (based on graphs from NT): Observation **(i)** matched expectations on the W and M data sets, whereas differences on the R and H data sets. For R, the difference is only subtle as the model's explanation to the why-question "Why did Kurt not Recover?" is not "Kurt did not Recover because of his bad Pre-condition, although he got Treatment." but "[...], which were bad although he got Treatment." which is on the second recursion in the reasoning process i.e., the treatment countering the state of condition and not affecting the condition itself. This difference becomes apparent in the graphical structure where the arrow from Pre-conditions to Treatment is inverted contrary to expectation. To illustrate one more drastic example using the data set of our Causal Hans example, here the discrepancy revolves around a totally different graph structure e.g. the learned model expects a direct cause-effect relation between age and mobility while also wrongly assuming that food habits have a detrimental effect on health. Therefore the answer to the question "Why is Hans's Mobility bad?" suddenly becomes "Hans's Mobility, in spite his high Age, is bad mostly because of his bad Health which is bad mostly due to his good Food Habits." which sounds very absurd. The ground truth SCM for this data set contains non-linear causal relationships, while NT makes linearity assumptions, which explains the wrongly learned graph structure. Observation **(ii)** only by looking at the CAT output, effectively using it as a graph distance or metric, we were able to tell that the learned model is very different from what we had initially expected. Put differently, it made apparent for the Causal Hans example that by simply adding an extra edge (here A, M) and flipping another (here N, H) we already get a big difference in what these graphs express/explain.

Does a CAT-based Regularization Penality Improve Graph Learning? While we have seen that CAT outputs are sensible explanations but that graph learning methods are still far from perfect in predicting graphs from data, in this final experiment we investigate how to use CATs *to improve* learning. Since CAT outputs contain (some) knowledge on causal relationships underlying the data, they should help in improving the overall prediction and sample efficiency of graph learn-

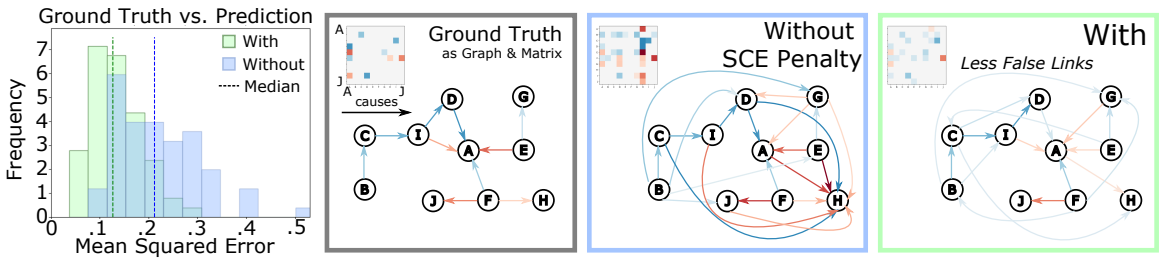

Figure 6: **Graph Learning Improves with Explanations.** Left: error distributions when performing graph learning with/-out CAT regularization (which is simply an added penalty term for inconsistent explanations), next to is the ground truth graph. Right (boxes): the predicted graphs, showing a decreased number of false positives. (Best viewed in color.)

ers. We take NT again as graph learner and add a simple regularization term to its loss that penalizes inconsistent explanations. We generate 70 random linear SCMs with respective observational distributions. Then we use graph learning to infer 70 more graphs, making 140 graphs in total. For each graph we generate 50 random why-questions to be answered, resulting in a data set of 7,000 explanations. All the details regarding this learning setup, such as for instance how to make make CAT differentiable for it to function as training signal, are being discussed in the Appendix. The graph learning is being performed in a data scarce setting with only 10 data samples per graph. Thus to infer the true causal structure the method ideally needs to perform sample-efficient learning. Fig.6 shows our results. The error distributions over all of the graphs are shown both with and without the CAT regularization. We also highlight the graph estimate upon which most improvement was observed. It can be observed that with the regularization the method can both identify more key structures while significantly reducing the number of false positives. For example many false links that pointed towards node H (like B to H or G to H) were removed while some key structures could now be recovered like the directed edge from node I to node A. While more experiments would be necessary to claim that indeed learning is (significantly) improved through explanations, our naïve learner already provides evidence in favor of our initial hypothesis.

Appendix B. Shortcomings of Previous Explainers (Short Case Study: CXPlain)

We take as an example for this section a popular explainer and vanilla CAT as introduced initially and provide a short comparative study/discussion. We start by reflecting on existing key ideas within the realm of explainable AI that makes use of causality. Since a great deal of existing literature is concerned with *causal attribution* we are going to discuss a representative case in the popular approach by Schwab and Karlen (2019) called CXPlain. Fig.7 top shows a why-question in a medical data setting. Particularly, patient Hans' medical condition is captured by different covariates (age, nutrition, health, mobility) and the question is concerned with why Hans' mobility value is lower on average than that of other patients. Fig.7 bottom then shows the answer given by CXPlain assigning three positive numerical scores to all variables except the one in question with the highest value being given to age, then nutrition and finally health. We can interpret this result as saying that all potential factors are actually being deemed relevant and "causal" to Hans' mobility. To briefly explain the setup: Hans' values were sampled from a synthetic SCM which CXPlain had access

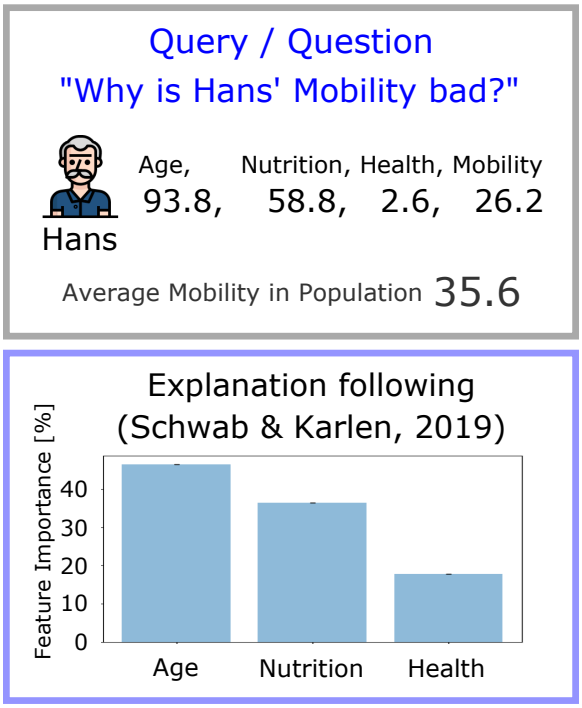

Figure 7: **Conceptual Limitations of Explaining Counterfactual Questions Without Use of SCM.** Refer to the text on the left for a discussion of the limitations. The figure shows a medical record for patient Hans and a question concerning the reasoning behind the sate of Hans' mobility. (Schwab and Karlen, 2019)'s method provides positive feature importance scores for the remaining variables. (Best viewed in color.)

to while training its surrogate explanation model. We trained 10 bootstrapped neural models using suitable parameters for the masking operation and loss function. Returning to our score distribution, this single observation makes apparent two important shortcomings of such causal attribution explainers: firstly, from the output we cannot deduce which is a direct (health in this case) and which are indirect (age and nutrition mediated via health) causes. Secondly, we have no information on the causal effect, that is, we cannot tell in which way a variable with high attribution will affect the predicted variable, for example the nutrition variable received a high importance score than age but age will have a detrimental effect on mobility whereas nutrition will have a beneficial effect. Two further, less important but still noteworthy, shortcomings are the following: thirdly, the attributions are deterministic. This might first be considered a feature, however, the causal mechanism of an SCM are only deterministic up to a realization of the exogenous variables. Therefore, we can have the exact same patient record for *different patients*. This cannot be captured by these previous attribution methods. Fourthly, when querying for random individuals we actually observe inconsistencies between the attributions themselves which is illogical since the patient records are being generated by the same causal mechanisms. For example in the Hans case we had age, nutrition and then health ordered from highest to lowest attribution. For a rather similar patient we observe that nutrition and age swap in importance. For yet another patient we observe that suddenly age and nutrition, which

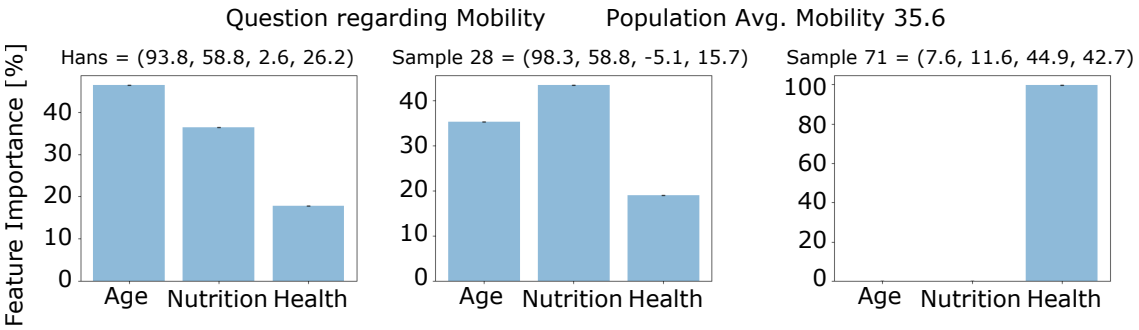

Figure 8: **Further Examples of CXPlain Shortcomings.**

previously played the most important role, are not important anymore. The reader is invited to look at the plots highlighting the last two deficiencies from Fig.8 of the appendix which shows the same setup as from Fig.7 but for different queries, that is, patients other than Hans. In conclusion, the four discussed deficiencies we pose as *desiderata* for our new approach to resolve. In summary, our approach should ideally: (Desideratum 1) differentiate direct from indirect causes, (D.2) capture qualtitative information on causal effects and (D.3) cope with stochasticity. CAT satisfies the desiderata through its SCM.

Appendix C. Multiple Noteworthy Short Discussions

The Importance of MM $\equiv$ SCM for CAT. While the MMC is a fundamental question that cuts to the core of human thinking and remains to be proven right or wrong (although we believe it to be true to the extent of representability through SCM), and while we used it to ultimately justify the usage of SCM to then derive the causal explanations we call CAT, still, to the actual existence and formalism of CAT the MMC's truth value is invariant. Put blantly, if the MMC were to be wrong, then the formalism of CAT and all proven properties remain *the same*. However, if MMC were to be true, then CAT in fact become a "stronger" formalism for causal explanations since they'd have a direct link to the MM. More importantly, one could make the case that they'd represent a "natural" formal pendant to the vague human explanations.

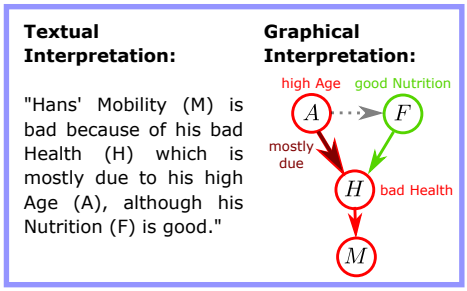

Figure 4: **Interpreting the CAT Output for Causal Hans.** The example illustrates all three rules. (Best viewed in color.)

Simpson's Paradox Example. Consider the well-known Simpson's paradox example for the medical setting of Kidney stone treatments from (Charig et al., 1986). The setting is given by T, K, R which are Treatment, Kidney Stone Size, and Recovery respectively, and further the graph is given by $T \to R, K \to \{T, R\}$. It is known that $T = 0$ denotes open surgery and $T = 1$ denotes Percutaneous nephrolithotomy (being a more involved procedure) and in the overall statistics for recovery of the patient (denoted by $R = 1$) we observe 78% versus 83% respectively, suggesting

that $T = 1$ is the better option. Yet, when looking at the confounder K values of patient recovery, we observe 93% versus 87% for a small kidney stone $K = 0$ and 73% versus 69% for a large kidney stone $K = 1$ respectively, suggesting that in fact $T = 0$ is better instead. This is the "paradoxical" situation, which is sensible from the *causal perspective*. If we now ask the single why-question for patient i with say values $T = 1, R = 0, K = 1$ on why i did not recover $r_i < \mu^R$ (where μ^R is the mean recovery of the data set), then we obtain a CAT output that reads as follows: *"Patient i did not recover because of the large kidney stone, although (s)he had Percutaneous nephrolithotomy."*

Hidden Confounders in Semi-Markovian Models. As we pointed out in the main text, CAT can naturally handle/extend to semi-Markovian models. For illustration, consider the non-Markovian alternative to the example from the paragraph above on Simpson's paradox, where K is a hidden confounder i.e., we only observe $T \to R$ as the graph. In a lot of practical settings we might at least be aware of the fact *that there is* hidden confounding present between the two variables and thus have an additional (dashed) bi-directed edge between T and R (case 1) and in the arguably worst case, said variable is fully undetected (case 2, in this case it is not necessarily a hidden confounder but simply a hidden cause, since we don't know if it is confounding or not—confounding meaning the same thing as *common cause*). Let's consider both cases, in case 1, the CAT output for the same question as before would read as: *"Patient i did not recover although (s)he had Percutaneous nephrolithotomy."* We note that simply the reasoning on K is not being delivered, naturally, since K is not in the SCM that the CAT process observes. For case 2, we'd observe the same reading due to the definition of the CAT construction. Here, however, we note that this case allows for a natural extension of CAT in which the reading could change to possibly, "Patient i did not recover because of *an unknown reason*, although (s)he had Percutaneous nephrolithotomy." Note that this semi-Markovian CAT now allows for reasoning with "unknown reasons" since the hidden cause K will certainly have a causal relation to R (since K is a cause) but the name of K will not be revealed (since K is hidden). With this example, we thus conclude that Markovianity can be leveraged by CAT.

A Well-behaved Algorithm for Generating CATs. The concepts of why-question, causal scenarios and *ER*i rulest hat we had to develop for the introduction of CAT algorithm come with several mathematical consequences which we now discuss. All of the subsequent results are simple and can be proven easily, still, their importance needs to be stressed since they make implications about the wide applicability of CAT.

Proposition 1 *For any causal scenario the rules ER*1 *and ER*2 *will be mutually exclusive.*

Proof First, we code the binary ordering relations $<, >$ to represent 0 and 1 respectively. Second, we observe that ER$i \in \{<, >\}, i \in \{1, 2\}$ always involves the triplet $T = (R(\alpha, 0), R(v_j, \mu_j), R(v_i, \mu_i))$. Third, let $\mathbb{T} := \{0, 1\}^3$ be the set of all such triples as their code words, so $T \in \mathbb{T}$. Looking at the total number of possible scenarios $|\mathbb{T}| = 2^3 = 8$, we easily see that ER1 covers codewords $\{010, 011, 100, 101, 000, 111\}$ and ER2 covers the codewords $\{001, 110\}$, and together they cover all codewords ER1 $\cup$ ER2 $= \mathbb{T}$. Since any single scenario $C_{i,j}$ is uniquely mapped to a codeword, it will either trigger ER1 or ER2 but never both. ■

Proposition 2 *The CAT recursion always terminates.*

Proof The recursion's base case is reached when a root node is reached i.e., a node i with $\mathrm{pa}_i = \emptyset$. An SCM implies a finite DAG, so root nodes are reached eventually. ■

Proposition 3 *The output of any causal structure learning algorithm can be used to compute CAT outputs.*

Proof The proof for this proposition is surprisingly simple in that the SCM $\mathcal{M}$ used in the CAT recursion is only required to provide some kind of numerical value α for the relation of any variable pair (V_i, V_j), that is, a matrix $\mathrm{A} \in \mathbb{R}^{|\mathbf{V}|\times|\mathbf{V}|}$ which represents a linear SCM or a SCM where each α represents a causal effect description. If the matrix A is an adjacency matrix living in $[0,1]^{|\mathbf{V}|\times|\mathbf{V}|}$, then we simply have no information about *ER*3 since all causal effects are assumed to be the same. Since any causal structure learning algoirthm will produce a causal graph represented by a matrix, we have that we can compute CAT outputs on it. ■

The beauty of Prop.3 can be fully appreciated when being put into the context of practical AI/ML research and application. It tells us that *any* causal graph learner ever invented and that will ever be invented can provide causal explanations on any query of interest consistent with the learned model thus reflecting the learnt. In practice this means that all prominent graph learning algorithms like NT (Zheng et al., 2018), CGNN (Goudet et al., 2018), DAG-GNN (Yu et al., 2019) and NCM (Ke et al., 2019) are all explainable[1]. On a concluding note to this section, we have a remark on SCM that allow for hidden confounder. CAT as presented Def.7 do not cover hidden confounders and we leave this for future work. However, we can always modify the algorithm to talk about "unknown reasons" when giving knowledge on $\mathbf{U}$.

Appendix D. CAT+Time, 2D Game Study: "Why did Mario go after the Goomba?"

It is entirely conceivable that our time-series Causal Hans example does not accurately reflect reality and that we can depict this more precisely. Theoretically, it could be that the relationship between Age and Nutrition during adolescence (Age $\leq$ 25) is primarily negative, which could be due to the consumption of alcohol or junk food, and subsequently (Age $>$ 25) turns positive again. In this scenario, two SCMs could be deemed valid based on a specific variable, thereby allowing for a seamless transition between the SCMs and the variables used within them. The intuition behind Gens. 1, 2 3, and Def. 9 would still be applicable in this case. It is crucial to employ the accurate causal graph and the appropriate data, as outlined in Def. 9 (see ϕ_1 and $\mathfrak{M}$). The Explanation Tree and the resulting explanations would continue to operate effectively, allowing for the Age variable to be accurately summarized in each instance. However, the explanation may evolve over time for some examples due to varying causal relationships.

We now turn our attention to agent behavior in a 2D grid-based game called CoinRunner (see Fig. 9), inspired by CoinRun (Cobbe et al., 2019). The game comprises various sprites, including Goldcoin, Powerup, Enemy (referred to as Goomba), Player (Mario or Luigi), and Goal. The player starts with 20 points and loses one point per second until they reach the Goal. Collecting Goldcoins grants 5 bonus points (BP), while colliding with the Enemy yields 9 BP if the Powerup has been collected beforehand, and results in a game over with a score of -20 if no Powerup was obtained.

1. The DAG learner in NT can be interpreted as a linear SCM but there is no guarantee.

The initialization and positioning of the sprites are independently randomized. Each game rollout R is represented by a sequence of frames f, each described by mostly binary variables.

The main difference from the Causal Hans example is that the focus is now on explaining an agent's behavior within a rollout rather than examining population statistics. Interestingly, it is intuitive that explaining behavior involves not only retrospective considerations but also anticipatory reasons and consequences of specific actions. To this end, Def. 9 has an anticipatory case added (Def. 11), which can be used to encode effects in the future within the current context when the future is still uncertain, or across contexts when the future has already occurred (e.g., "Colliding with the Enemy positively impacts defeating the Enemy at the time step."). The abstraction of the CAT+Time defs. to the binary case is rather direct and does not require much elaboration. Further details are provided in the appendix.

For this game, we can define several deterministic agent behaviors, such as Coincollector, Killer, or Optimal. For the purpose of demonstration, we will highlight the Killer agent behavior. This behavior primarily focuses on collecting the Powerup and colliding with the Enemy when both Powerup and Enemy are present, before proceeding to the Goal. Similar to the dynamic Causal Hans example, various stationary processes can be identified from this. These primarily depend on the game's state (variables of a frame). For our Killer agent, processes can be most easily described based on the existence of sprites. We call C a context, which brings us to the currently valid SCM and thus to a stationary subprocess. Specifically, for the Killer agent, we have defined the following three subprocesses: (i) $C_{K,1}$ = powerup and opponent exist', (ii) $C_{K,2}$ = powerup does not exist and opponent exists', and (iii) $C_{K,3}$ = 'neither exists'. For this purpose, we have implemented an imperfect Killer agent and recorded 500 rollouts frame by frame. By imperfect, we mean that, with very low probability, it can also exhibit other behavior. Together with a bit more noise, we then used Lasso, VARLiNGAM (Hyvärinen et al., 2010), and Granger on conditioned frame sections, depending on our contexts C, to generate graphs that we want to assume as causal for the moment. The learned graphs, a description of the methods used, and further CAT parameters are not essential for the main part and the demonstration of the application of this work and have also been moved to the appendix.

As a result, we can already causally explain questions like "Why did Mario jump on the Goomba?", "Why is Mario targeting the Goomba?" or "Why does Mario run into the Goal?", if the question is valid in the specific time step and, as in our case, at runtime. For this purpose, the current frame is used to identify the currently valid SCM, and the rest of the recorded dataset serves as the explanation basis. Fig. 9 shows a scenario in which Mario is running towards the opponent with the retrospective explanation on Fig. 9 (right). The corresponding *anticipatory* explanation is:

> Targeting the enemy has a positive effect on targeting the enemy, the existence of the enemy, colliding with the enemy and a negative effect on the score, targeting the goal and killing the enemy in the next time step.

The CoinRunner example has contributed only a relatively small proportion to the main body of this work. We intend to use this appendix to provide more detailed information about the game, the learned causal graphs, and additional explanatory examples, as well as to discuss the minor definition modifications required for the implementation of causal explanations.

Fig. 10 displays the scope of the implemented game as described in the main body of this work. On the left side is the playing field with all possible individual sprites. The agent is currently (still)

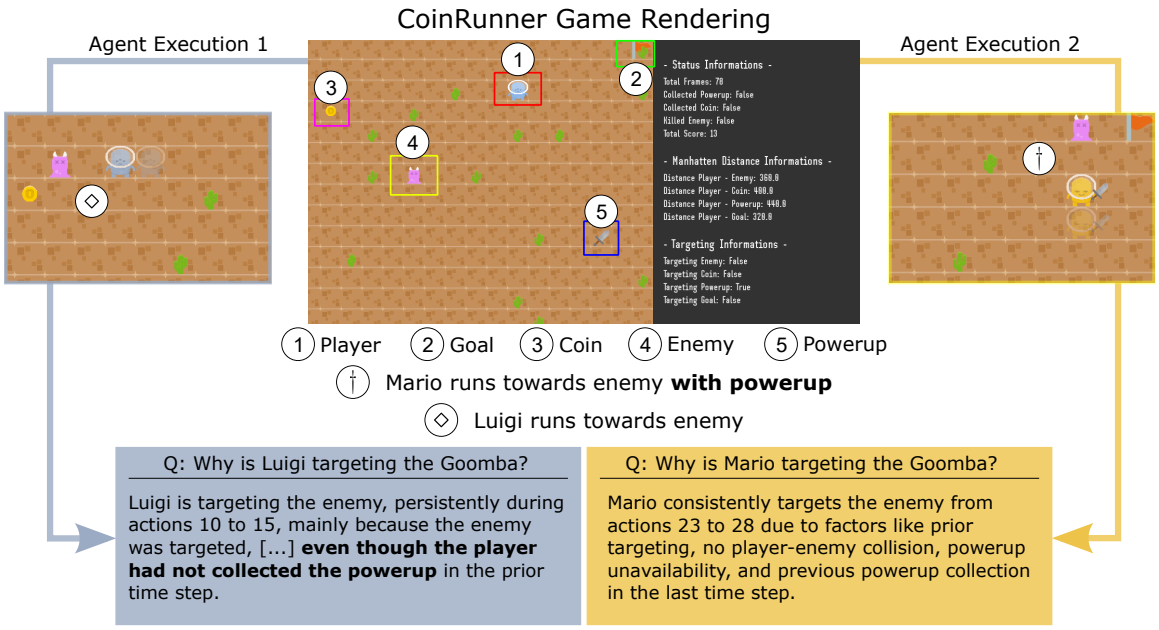

Figure 9: **"Why Was That Happening?"** Causal explanation for the behavior of two agents (referred to as Luigi and Mario) from the same player type in different game situations of the 2D CoinRunner game using the generalized CAT framework. (Best viewed in color)

represented as a blue avatar. The yellow version of the avatar indicates that the power-up has been collected. On the right side is an information board, which is not of further interest to us in the context of this work.

To provide valid questions and meaningful encoding of causal relationships, we will first adapt the definitions of the CAT+Time for our use case. Key differences include that we are no longer interested in population statistics, but rather in the agent's behavior during a rollout. Anticipatory explanations were already discussed in the main body, and a definition was provided. *At runtime* of a rollout as in our case this is limited to a pronounciation like 'has a positive' or 'has a negative effect', affecting the *EI* function in Def. 9 to exclude *ER* and solely rely on $s(\alpha_{X \to Y})$. In the general case, when the entire time series is available, further explanations can also be generated in the future direction. For the retrospective case the definition remains unchanged. Further, we have to modify the valid Why Question, the Causal Scenario, and the Fundamental Rules, which were introduced for the Causal Hans example. It should also be mentioned that we need to adjust the rules here partly because our game randomizes the sprites independently on the playing field. For this reason, individual rollouts vary in length regardless of the agent's behavior. In cases where rollouts are more comparable, population statistics can also be used.

Definition 12 (Why Question - Behaviour) *Let F_t be a frame of a rollout R, and let $x_t \in Val(X_t)$ be an instance of $X_t \in \mathbf{V}$ of the contextual-dependend SCM $\mathcal{M}$. We define $Q_{M,X_t} := x_t$ as a valid question if X_t is a binary variable, M is the appropriate contextual SCM for F_t, and Q is true.*

Definition 13 (Causal Scenario - Behaviour) *The tuple $C_{XY} := (\alpha_{X \to Y}, x_t, y_k)$ is referred to as a behavioral causal scenario.*

Figure 10: **Basic CoinRunner game board including information board on the right side.** Agent, gold coin and powerup have been initialized and are placed at random positions in the environment. The target sprite is placed at one of the four corners. Best viewed in colors.

Definition 14 (Fundamental Rules - Behaviour) *Let C_{XY} denote a causal scenario, let $s(x) \in \{-1, 1\}$ be the sign of a scalar. We define FOL-based rule functions as*

(ER1) If $x \in \{0, 1\}$, then: $(s(\alpha_{X\to Y}) < 0 \wedge x_t \oplus y_k) \vee (s(\alpha_{X\to Y}) > 0 \wedge x_t == y_k)$

(ER2) If $x \in \mathbb{R}$, then: $s(\alpha_{X\to Y})$

indicating for each rule ER$i(\cdot) \in \{-1, 0, 1\}$ how the causal relation $X\to Y$ satisfies that rule.

We previously described the three sub-processes for the Killer agent: (i) $C_{K,1}$ = 'powerup and opponent exist', (ii) $C_{K,2}$ ='powerup does not exist and opponent exists', and (iii) $C_{K,3}$ ='neither exists'. In Fig. 10, for example, all sprites are present. The Killer agent would deterministically move towards the power-up since both the power-up and opponent exist. Afterward, a new subprocess directs the agent to move towards the opponent.

To obtain causal graphs for each context C, we first conditioned the rollouts from 500 runs of a nearly deterministic agent tailored to Killer behavior on the respective contexts. Here we also incorporated Margin Frames and a small amount of noise. For each of the 500 rollouts, we generated causal graphs using three methods: (i) Granger causality combined with Vector Autoregression (VAR), (ii) VARLiNGAM, and (iii) Lasso. The resulting Graphs for each method and context were individually averaged at the end.

For all methods, we considered a time lag of 1 and excluded instantaneous effects. First, we trained a VAR model for method (i) and conducted a Granger causality test for variables with p-values less than 0.05. We included the coefficients of the VAR model in the causal graphs when the significance level of the granger test was also less than 0.05. For method (ii), we ran the VARLiNGAM model without making significant changes to the default parameters. As for method (iii), we used a LassoCV implementation with 5-fold cross-validation.

Fig. 11 displays the corresponding results. Rows with the prefix '-1_' have a causal effect on columns with the prefix '0_'. The variables used in the graphs are mostly self-explanatory and binary to provide meaningful explanations.

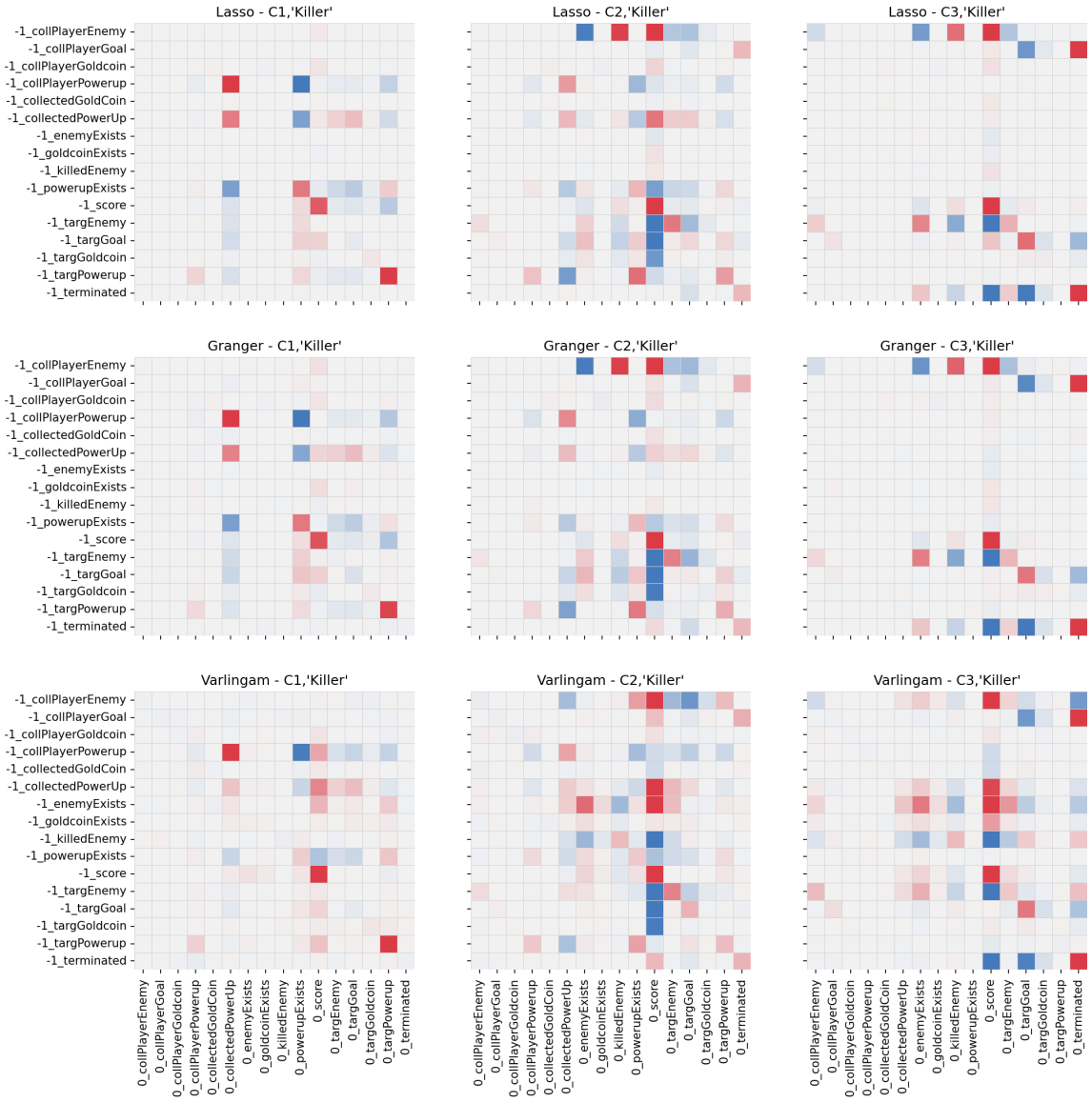

Figure 11: **Comparative Causal Diagrams for Behaviour Type Killer Using Granger, Lasso, and VARLiNGAM for $C_{K,1}$, $C_{K,2}$, $C_{K,3}$.** Red indicates positive influences, while blue represents negative effects.

It is important to note that the process of killing or picking up a sprite is divided into several steps in the time series. First, an object is "targeted," then "collided," and finally, it is no longer present or has been picked up in the next timestep. Additionally, we employed cosine similarity

for target tracking, which offers an advantage over Manhattan Distance and Euclidean Norm, as it allows targeting only one sprite at a time.

Using the CAT+Time approach, we can ask questions about why an agent performs a specific action in a given run if we consider the graphs to be causal. For example, in the main section, we answered the question, "Why is Mario targeting the Goomba?". Fig. 9 also shows what happens when Luigi, with the same behavior type, mistakenly targets the Goomba. Additionally, we compiled a collection of different answers from the three models to the questions: "Why is Mario targeting the Goomba?" (Tab. 2), "Why did Mario jump on the Goomba?" (Tab. 3), and "Why did Mario run into the goal?" (Tab. 4) in the corresponding tables. We used a uniform expression of encoding to maintain comparability, and the response width for both the retrospective and anticipatory parts was uniformly limited. Anticipatory effects are limited to one time step ahead.

Why did Mario target the Goomba?	
Model	**Explanation**
Lasso	***Retrospective:*** Mario is targeting the enemy, constantly over action(s) 23 to 28, mostly because the enemy was targeted, because the player did not collide with the enemy, because the powerup did not exist and because the player had already collected the powerup continuously in the previous time step. ***Anticipative:*** Targeting the enemy now has a positive effect on targeting the enemy, the existence of the enemy, colliding with the enemy and a negative effect on the score, targeting the goal and killing the enemy in the next time step.
Varlingam	***Retrospective:*** Mario is targeting the enemy, constantly over action(s) 23 to 28, because the enemy was targeted, because the player did not collide with the enemy and because the enemy did exist continuously in the previous time step. ***Anticipative:*** Targeting the enemy has a positive effect on targeting the enemy, colliding with the enemy, collecting the powerup, the existence of the enemy and a negative effect on the score and targeting the goal in the next time step.
GrangerVAR	***Retrospective:*** Mario is targeting the enemy, constantly over action(s) 23 to 28, mostly because the enemy was targeted, because the player did not collide with the enemy, because the player had already collected the powerup and because the powerup did not exist continuously in the previous time step. ***Anticipative:*** Targeting the enemy has a positive effect on targeting the enemy, the existence of the enemy and a negative effect on the score, targeting the goal, killing the enemy and targeting the goldcoin in the next time step.

Table 2: Causal explanation for the question 'Why did Mario target the Goomba?' during a recorded rollout.

Why did Mario jump on the Goomba?	
Model	**Explanation**
Lasso	***Retrospective:*** Mario is colliding with the enemy because the enemy was targeted (0.056) due to the action before. ***Anticipative:*** Colliding with the enemy has a positive effect on the score (4.338), killing the enemy (0.483) and a negative effect on the existence of the enemy (-0.479), targeting the goal (-0.197), targeting the enemy (-0.182) and targeting the goldcoin (-0.029) in the next time step.
Varlingam	***Retrospective:*** Mario is colliding with the enemy mostly because the enemy was targeted (0.059), because the player did not collide with the enemy (-0.02), because the enemy did exist (0.015) and although the player did not collide with the goldcoin (0.014) through the action before. ***Anticipative:*** Colliding with the enemy has a positive effect on the score (4.331), the existence of the powerup (0.23), targeting the powerup (0.15) and a negative effect on targeting the goal (-0.38), targeting the enemy (-0.235) and collecting the powerup (-0.219) in the next time step.
GrangerVAR	***Retrospective:*** Mario is colliding with the enemy because the enemy was targeted (0.04) due to the action before. ***Anticipative:*** Colliding with the enemy has a positive effect on the score (4.458), killing the enemy (0.486) and a negative effect on the existence of the enemy (-0.486), targeting the goal (-0.24), targeting the enemy (-0.201) and targeting the goldcoin (-0.04) in the next time step.

Table 3: Causal explanation for the question 'Why did Mario jump on the Goomba?' during a recorded rollout.

Appendix E. Details for Human Study

We instructed $N = 22$ participants to answer our questionnaire (see appendix Fig.12). The questionnaire asked the following questions: *"Given a pair of variables, does a causal relationship exist (existence)? If yes, then which is the cause and which is the effect (direction)? If there are multiple causes for a single variable, then how impactful is each of the causes (preference)?"* All of these questions, alongside their responses, are of *qualitative and subjective* nature. It is important to note that the participants *do not* perform the actual induction from specific, provided data like the algorithms do i.e., the human subjects are not given the variable names nor concrete data points that would allow them to find the rules for the specific data sets. Instead, they were only given the variable names/depictions, thereby having to induct from personal experience/understanding essentially. This approach to human induction is related to the experimental setups in (Griffiths and Tenenbaum, 2006; Hattori, 2016).

The motivating lead research questions we intended to answer, and in fact do answer successfully with this experiment, are: What are SCM that (some) human could model? How does overlap

Why did Mario run into the Goal?	
Model	**Explanation**
Lasso	***Retrospective:*** Mario is colliding with the goal because the goal was targeted (0.044) with the action before. ***Anticipative:*** Colliding with the goal has a positive effect on termination of the game (0.518), the score (0.029) and a negative effect on targeting the goal (-0.377) and targeting the goldcoin (-0.041) in the next time step.
Varlingam	***Retrospective:*** Mario is colliding with the goal mostly because the goal was targeted (0.055), because the player had killed the enemy (0.017), because the game is not terminated (-0.011) and although the player had already collected the powerup (-0.015) the time step before. ***Anticipative:*** Colliding with the goal now has a positive effect on termination of the game (0.505), killing the enemy (0.012) and a negative effect on targeting the goal (-0.374), targeting the goldcoin (-0.055), the existence of the enemy (-0.018) and targeting the enemy (-0.014) in the next time step.
GrangerVAR	***Retrospective:*** Mario is colliding with the goal mostly because the goal was targeted (0.023) and because the goldcoin did exist (0.01) the time step before. ***Anticipative:*** Colliding with the goal now has a positive effect on termination of the game (0.577), the score (0.026) and a negative effect on targeting the goal (-0.434) and targeting the goldcoin (-0.054) in the next time step.

Table 4: Causal explanation for the question 'Why did Mario run into the Goal?' during a recorded rollout.

for human-based SCM occur? How do subsequent CATs (Def.7) between humans and algorithms differ? In a nutshell, we wanted to investigate the similarity of SCMs between subjects in addition to the similarity between subjects- and algorithm-based CATs.

A caveat regarding the analysis and explanation of human judgements is that sample bias may distort conclusions. Sample bias has long been identified within the behavioral and social sciences as limiting the generalization of results obtained in a specific sample to the population. A common methodological fix to counteract such biases is to increase the sample size, see (Daniel, 2017) for a recent application and discussion. Certainly, the observed sample will affect the way the difference (to e.g. algorithm-based CAT) turns out to be, but then again our research question is *not* concerned with all possible human explanations, but any. Furthermore, we chose data sets that model very general examples and thus offer accessibility to the general population since no single person might be an expert. Ultimately, this way of designing our experiment, while not removing sample bias of course, renders the bias's qualitative effect onto our subsequent investigation negligible.

In the following we provide a discussion of several interesting and important insights discovered through the human user study. Nonetheless, it is important to note that our results like most

modern day interpretations of human behavior are of conjectural nature – sensible, educated guesses essentially. During this discussion, we will point to specific aspects of the descriptive statistics displayed in appendix Fig.13. The questionnaire contains four examples with two, three, three, and four variables (or concepts) respectively that are being visually depicted in addition to a concise textual description. We randomized the textual description of up to three variables across all examples for any randomly selected participant. Doing so, we allow for the randomized concept to reverse causal influence directions, thus, diminishing the bias of chance-selecting said causal direction – in a nutshell, this randomization scheme helps us in controlling for explanation variance (or leeway) of the subjects. Nonetheless, we still observed that for any variable pair (X, Y) the meanings of X and Y themselves could be interpreted differently, which ultimately resulted in False Negatives regarding agreement i.e., people will disagree technically although they actually agree. To give a concrete example, consider the following: pre-condition in Example 2 can be interpreted as "the length of the medical history of a patient" (negative; increasing implies lower chance of recovery) opposed to "the state of well-being of a patient" (positive; increasing implies higher chance of recovery), thereby some subjects might choose $Z_1 \to R$ while others will choose $Z_2 \leftarrow R$ where Z_i are the different explanations of the "pre-condition" concept (and R denotes recovery), yet all subjects agree on an existing relation between the two variables: $Z_i \leftrightarrow R$. Also, some variables/concepts were more stable in their explanation variance. To give yet another specific example, altitude and temperature in Example 1 (appendix Fig.12) are stable concepts while the aforementioned pre-condition in Example 2 is unstable (due to its explanation variance/leeway). More importantly these different explanations due to the ambiguity inherent in language become visible within the statistics. To stay inline with the previous example, consider the medical example within appendix Fig.13 (second row, middle) and specifically consider the edges $T \to R$ and $Z \to R$. For the former relation the agreement between subjects is evident i.e., the majority of human subjects will select this edge. For the latter relation, we clearly see the two previously discussed explanations that subjects employ during edge decision. I.e., for some subjects the edge between Z and R is positive and for some others it is negative, while naturally all agree upon there being a relation between the variable pair $(Z \leftrightarrow R)$ opposed to there being no relation $(Z \not\leftrightarrow R)$.

We observe a systematic approach and thereby non-random approach to edge-/structure-selection by the human operators, see any of the subplots within appendix Fig.13. Furthermore, there are only a few clusters even with increasing hypothesis space. Both the systematic manner and the tendency to common ground are evidence in support of the MMC hypothesis (MM $\equiv$ SCM, Hyp.1) and its implied argument on "true" SCM information reachable from the overlapping MM-based SCMs or SCMs.

Although we randomize the order of variables in addition to consistently presenting them in a simple line with the intention of not inducing any specific sorting/structure to avoid bias, we still observed apparent, unintended subject behavior. For instance, subject number 5 only considered pairs presented next to each other as being questioned although the other combinations are meant to be queried as well. While additional research needs to corroborate these observations, our data suggests that attention might have decreased over the course of the experiment for a subset of subjects as suggested by e.g. subject number 7 where overall agreement with the subject majority is to be found but eventually at the very last example "mistakes" occur (specifically, the subject highlighted that "increasing age increases mobility", in stark disagreement with the majority of participants). We also observe that the increase in hypothesis/search space (i.e., more variables) comes with an increase in variance. This variance increase can be argued to be due to the progressive difficulty

of inference problems as well as decreased levels of attention and potential fatigue across the duration of the experiment (e.g. consider the duplicate plots, third column, in appendix Fig.13 where the number of unique structures that are being identified increases significantly). Yet another interesting observation concerns the aspect of time, consider subject number 17 where there is a cycle between treatment and recovery where the subject likely thought in terms of "increasing treatment increases speed of recovery *which subsequently* feeds back into a decrease of treatment (since the individual is better off than before)" which seems like a valid inference but clearly considers the arrow of time. Yet another observation, some subjects faced questions of variable scope e.g. if there is a causal connection between food habits and mobility, then some subjects considered energy as the mediator and since energy is not part of the variable scope, confusion might arise whether to place an edge between food habits and mobility or not. In fact, for such a scenario the correct answer is to place an edge, since there exists a causal path from food habits to mobility, via energy, even if energy is not displayed. I.e., in causality, an edge can/will talk implicitly about all the more fine-grained variables that are part of the causal edge/path.

The second data set is an instance of the famous Kidney Stone example (Peters et al., 2017), where Z is a confounder that indicates the pre-conditions in terms of e.g. the size of the kidney stone, and it also illustrates the famous Simpson paradox (Simpson, 1951; Pearl, 2009; Peters et al., 2017) where the recovery will favor one treatment in the overall statistics while being better for all of the non-consolidated views for the other treatment. We observe that not a single subject places the edge pre-condition to treatment ($Z \to T$) which is arguably at the core of Simpson's paradox. This observation gives an additional cue on why the phenomenon is called paradox because no human subject expects the existence of this connection and even actively neglect the existence.

We observe that the human-based CATs match the Ground Truth CATs *perfectly* up to the R data set CATs, which is also the "Result" in Fig.5 i.e., the "Mode" approach returns the correct CATs while the "Greedy" approach chooses the wrong edge type for Z and R. After further investigation, we believe to have found several explanations for this "human" mistake that we discuss extensively in the appendix. On another note, we observe that the overall flawless performance of human-based CATs speaks for superiority over algorithmic graph learner-based CATs. To conclude this paragraph, let us appreciate one such drastic difference in explanations, which in fact occurred on our lead example "Causal Hans":

Humans: *"Hans's Mobility is bad because of his bad Health which is mostly due to his high Age, although his Food Habits are good."*
Causal Discovery Algorithms: *"Hans's Mobility, in spite his high Age, is bad mostly because of his bad Health which is bad mostly due to his good Food Habits."*

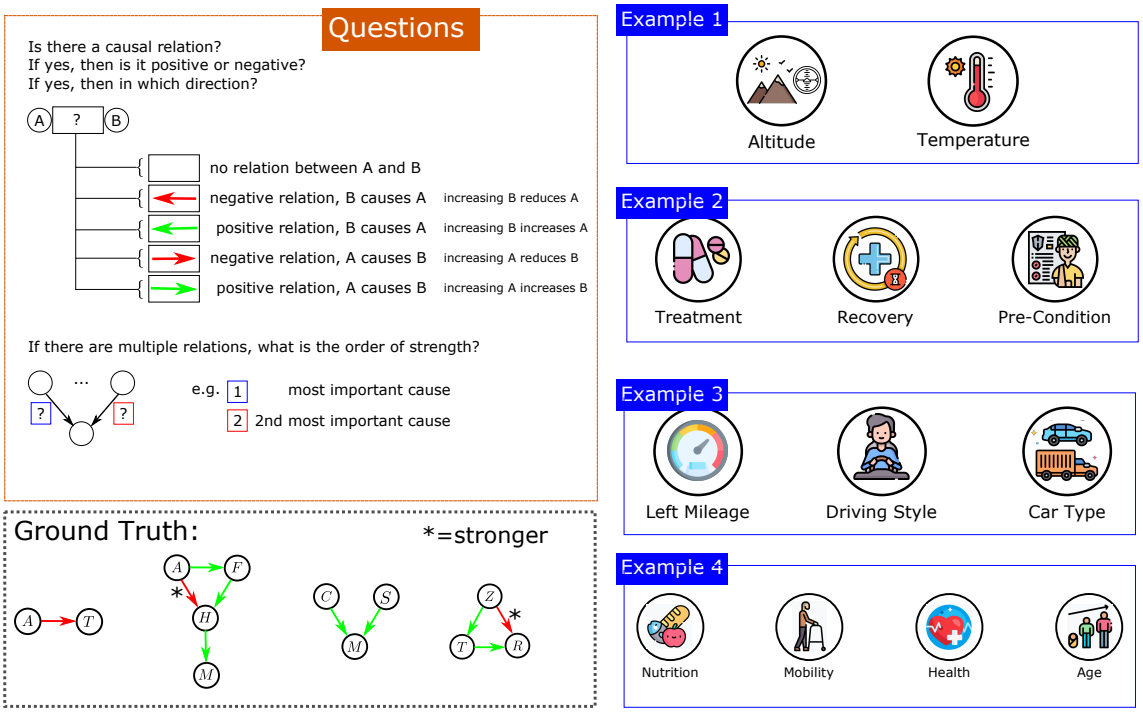

Figure 12: **Experiment Setup for the Human Case Study.** The participants are being asked two questions: whether there is a directed relation between some variable pair A and B, and when there are multiple causes how they behave relatively i.e., the order of strength in relations. We avoid bias in drawing relations by randomizing the order and presenting the variables in a sequence. Induction is being performed from personal "data"/experience, rather than by looking at a matrix of data points. To avoid bias in drawing relations, we don't provide any hints on a graph structure and we randomize the sorting of the variables. To provide more clarity we depict the names of the concepts with additional illustrations. The participants are asked to perform induction based on personal data/experience i.e., they only see the orange and blue boxes. (Best viewed in color.)

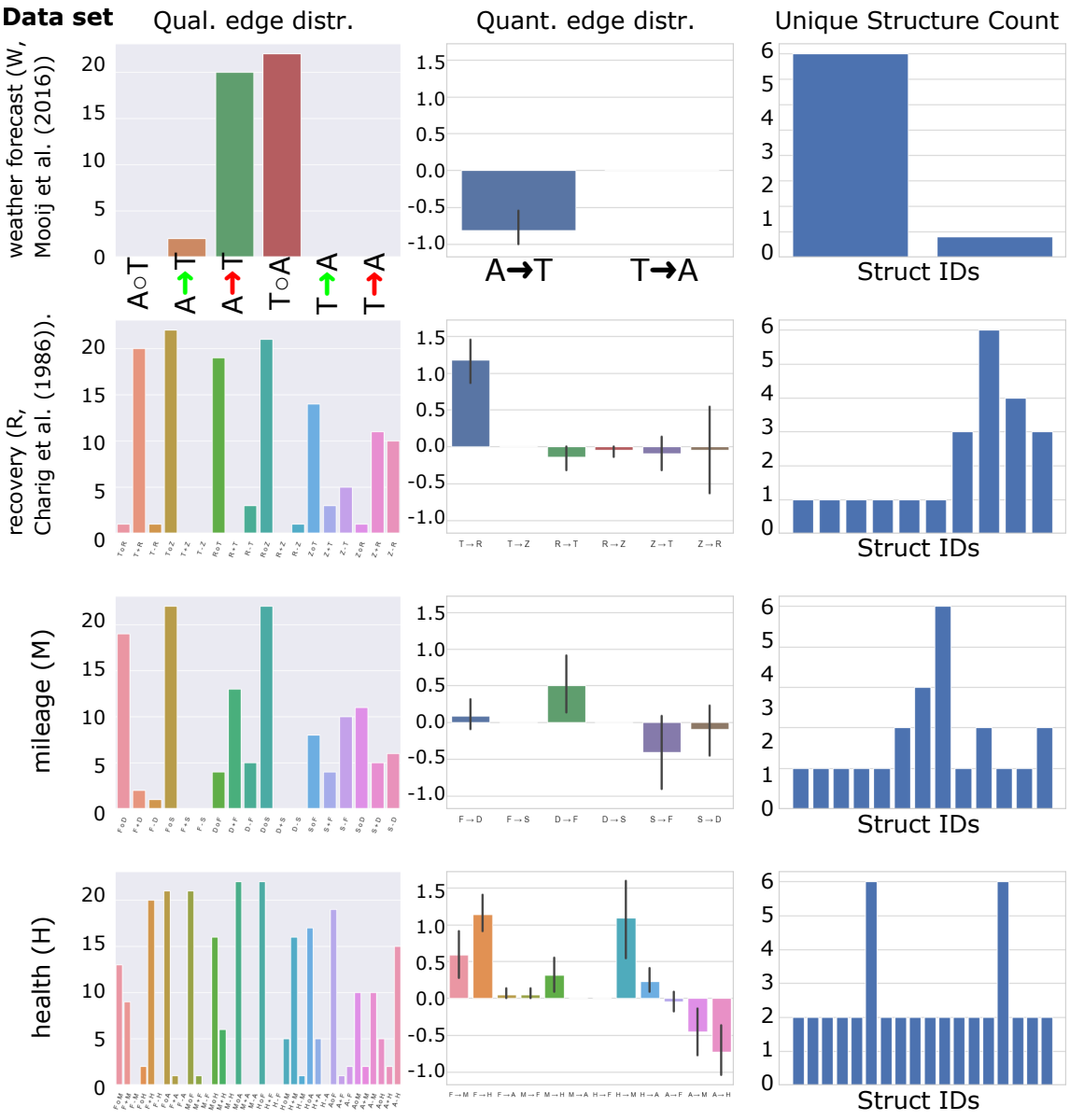

Figure 13: **Human Data Analysis: Qualitative, Quantitative, and Uniqueness.** Statistics collected from the human data ($N = 22$). The rows denote the four data sets: weather forecast (W, Mooij et al. (2016)), recovery (R, Charig et al. (1986)), mileage (M), and our synthetic health data set (H). The columns: qualitative edge distributions that show for each of the different edge type how often it was chosen respectively (left). Qualitative because it incorporates ACEs. A green edge denotes positive ACE, whereas red means negative ACE (i.e., increasing X would decrease Y). Quantitative edge distribution for each edge where the error bars denote confidence intervals (middle), and the unique structure counts where each bar depicts the frequency of a qualitative structure discovered by the human subjects (right). Extensive elaboration on the setup, execution and results of this human study are to be found in the corresponding appendix section. (Best viewed in color. Since the plots get dense with increasing combinations of variable pairs, please consider zooming in to read the labels for any detailed view of the results.)

Appendix F. "Food for Thought" for Future Work: A Hypothesis on the Representational Suitability of SCMs as Mental Models

As a question of cognitive science and psychology, we have placed this section in the appendix for the interested reader. Our Hypothesis revolves around the idea that SCMs are a suitable representation of human mental models. It has been argued that at the core of a human mental model (abbreviated MM in the following) the illustration of one's thought process (regarding the understanding of world dynamics) is to be found (Simon, 1961; Nersessian, 1992; Chakraborti et al., 2017). The difficulty of said thought process illustration is partly due to circular and abstract terms like explanation and interpretations for which we do not provide an explicit definition as this is up to philosophical debates and ideally we keep the idea more general than what has been done previously in explainable AI/ML where "explanation equals pixel attributions" in many cases. Assuming the world dynamics to be governed by causality we observe that humans are capable of modelling both causal relationships between endogenous variables and additionally information on the strength of said relationship. Put differently, MM model a causal graph and corresponding causal effects akin to the formal notions from the previous section. Consider the following real world example:

MM Example. *At any given time a human has a state of overall health (relating to fat-muscle ratio, allergies and diseases, etc.) and mobility (relating to the general freedom and flexibility of movement, e.g., a gymnast is more mobile than the average person). Now, the MM allows inferring (1) that mobility is being (partially) caused by something else (for instance health, e.g., being overweight decreases one's mobility) and (2) that different events can have different "strength" e.g., that an average car accident causes more harm to the individual's mobility than an average workout session causes good.*

A natural candidate for capturing the two properties from the MM example formally are SCM, thereby we hypothesize the following:

Hypothesis 1 (MM Conversion, short MMC) *The parts of the MM that are being used for encoding the causal relationships of the variables of interest can be formally captured by a corresponding SCM, in short this "equivalence" can be denoted as MM $\equiv$ SCM.*

While the MMC hypothesis leaves room for notions not captured by mathematical rigor, it suggests an equivalence to SCM regarding the causal aspects. The MM example has motivated the MMC hypothesis which itself suggests *a justification of using SCM in the first place*.

Implications of MM $\equiv$ SCM. If we accept that MM $\equiv$ SCM, then we can use SCMs as an adequate proxy to the MM. Furthermore, any useful property of SCM implies corresponding aspects back in the MM. We immediately observe one such key property of SCM namely comparability. That is, if one is given say two different SCMs that are defined over the same endogenous and exogenous variables (so only differing in the actual parameterizations) then one can compare said SCMs i.e., there exists a notion of distance. For the linear case, we can easily prove this by construcing an example metric space.

Definition 15 *We define a function $d(\mathcal{M}_1, \mathcal{M}_2) = \sum_{i \neq j} |\mathcal{M}_1(j,i) - \mathcal{M}_2(j,i)| + q(P_1, P_2)$ where q is the square-root of the Jensen-Shannon Divergence (JSD), $\mathcal{M}_k = \langle \mathbf{U}_k, \mathbf{V}_k, \mathcal{F}_k, P_k(\mathbf{U}_i) \rangle$ for $k \in \{1,2\}$ such that $\mathbf{V}_1 = \mathbf{V}_2$, $\mathbf{U}_1 = \mathbf{U}_2$, $\mathcal{F}_k$ define linear functions in $\mathbb{R}$, and in slight abuse of notation $\mathcal{M}_k(j,i)$ is the causal effect α from V_j to V_i.*

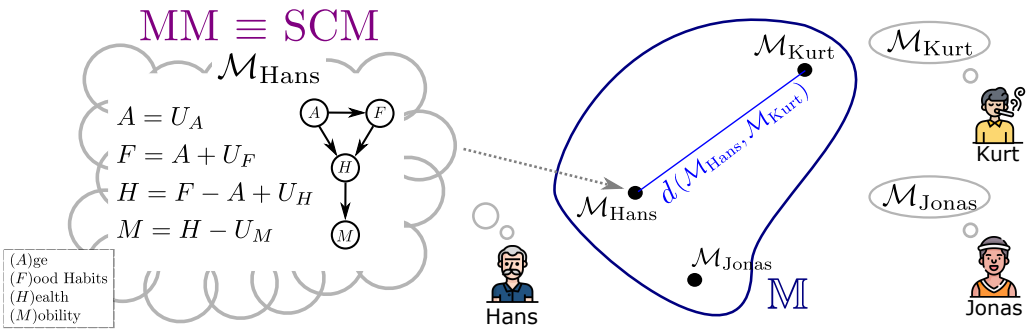

Figure 14: **MMC Hypothesis and Linear SCM Metric Space.** Left: Accepting Hyp.1 means that the MM of Hans is an SCM. Right: Different linear SCM (from different individuals) can be compared, an example metric space for $(\mathbb{M}, d)$ is given by Prop.4. (Best viewed in color.)

Proposition 4 *Let d be as in Def.15 and let $\mathbb{M}$ denote the set of all linear SCM defined over the same exogenous and endogenous variables, $\mathbf{U}, \mathbf{V}$. Then $(\mathbb{M}, d)$ is a metric space.*

Proof The absolute difference on the real numbers is a metric (i.e., positive-definiteness, symmetry, and triangle-inequality hold) therefore holding for the "dependency" terms from $\mathcal{F}$. Furthermore, q was chosen as the Jensen-Shannon-Metric. Finally, metrics are closed under summation. ■

Prop.4 is just one example of what might be considered a sensible metric space for a subset of all SCMs. What it does is compare each of the linear coefficients for any causally related tuple of variables, aggregating the sum, and further adding a divergence term between the defined distributions over the exogenous variables. This comparability and the visual intuition behind MMC are illustrated in Fig.14. We now state our **first key observation** following Hyp.1 and Prop.4: *the existence of a "true" SCM is in fact justified i.e., there exists an underlying data generating process for any data and the MM of any person might or might not coincide with that SCM.*

On another note, consider the fact that while the "true" SCM represents the concept of objectiveness, oppositely, the MMs are of subjective nature (that is, every human has their own subjective life experience). Coming back to MM ≡ SCM, we see that Prop.4 further implies that MMs are also capable of dis-/agreeing with each other. With this at hand, we now state our *second key observation*: in most practical cases having access to many SCM-encodings of subjective MMs can ultimately lead in their overlap-agreement to (parts of) the objective "true" SCM. There is certainly no guarantee since all available MM-SCM samples can in fact be wrong, however, note the emphasis on *in most practical cases*—therefore, identifying this overlap in MM (or SCM) for a specific problem is highly valuable for AI/ML research.

Our final, *third key observation* is concerned with explanations. Existing literature views explanations as *derivable* from MMs and thus implicitly containing some information on the MM (Chakraborti et al., 2017) and since MM ≡ SCM, we argue that there must exist an equivalent of the human notion of explanation *within SCM*. This justifies our further investigation on SCM-based explanations, which eventually leads to the formalism of CATs. The benefits of an approach using explanations derived from SCM are two-fold (1) that by construction they are human understand-

able allowing for explainable ML in which models can reason about the learnt and (2) that the models themselves become better, as they need to account for consistency in explanations, which is beneficial to any downstream-task.