Causal Abstraction with Soft Interventions

Author names withheld

Editor: Under Review for CLeaR 2023

Abstract

Causal abstraction provides a theory describing how several causal models can represent the same system at different levels of detail. Existing theoretical proposals limit the analysis of abstract models to "hard" interventions fixing causal variables to be constant values. In this work, we extend causal abstraction to "soft" interventions, which assign possibly non-constant functions to variables without adding new causal connections. Specifically, we (i) generalize τ-abstraction from Beckers and Halpern (2019) to soft interventions, (ii) we propose a further definition of soft abstraction to ensure a unique map ω between soft interventions, and (iii) we prove that our constructive definition of soft abstraction guarantees the intervention map ω has a specific and necessary explicit form.

Keywords: Causal Abstraction, Soft Interventions, Structural Causal Models

1. Introduction

Causal modeling is a crucial tool to represent, reason, and act on systems composed of causally connected independent mechanisms (Peters et al., 2017; Bareinboim et al., 2022). A causal analysis requires a choice of variables; that is, we must choose the level of *abstraction* in which to couch an analysis (Woodward, 2015). The question becomes especially salient with a large number of densely connected variables that are difficult to analyze or understand. As such, the fundamental concept of *causal abstraction* has played a key role in the scientific investigation of complex phenomena, including weather patterns (Chalupka et al., 2016), brains (Dubois et al., 2020a,b), and deep learning artificial intelligence models (Chalupka et al., 2015; Geiger et al., 2020, 2021).

Within causal modeling, there is a recent growing body of literature on formal theories of causal abstraction (Chalupka et al., 2017; Rubenstein et al., 2017; Zennaro, 2022). However, following the widespread do-operator (Pearl, 2009), existing theories of causal abstraction only consider *hard interventions*, where a variable is fixed to a single value. Not all mechanistic changes to a real-world system can be represented by a hard intervention. Increasing network weights in a deep learning model is not a matter of fixing the values in representations; the causal mechanisms have been altered, but not fixed. In this context, soft interventions formalize the replacement of causal mechanisms with different and possibly non-constant mechanisms (Eberhardt and Scheines, 2007).

In this paper, we extend previous work on τ-abstraction (Beckers and Halpern, 2019) to address the description of non-constant abstracted mechanisms. In Section 5, we introduce a generalization of the underlying formal concepts and we prove that, while correctly handling low-level soft interventions, the resulting definition of τ-abstraction is ambiguous for high-level soft interventions. Consequently, we propose a novel definition of soft abstraction by strengthening the consistency requirement of τ-abstraction on endogenous settings. Notably, we are able to prove that our extension maintains desirable properties of τ-abstraction and that is equivalent to it whenever the two models are hard-intervened. Finally, in Section 6, we specialize our definition to the constructive case, where each high-level variable depends on a subset of the low-level variables, as in Figure 1. As a further contribution, we prove that constructive soft abstraction implies a unique explicit form on the function mapping interventions from the lower- to the higher-level model.

© 2023 .

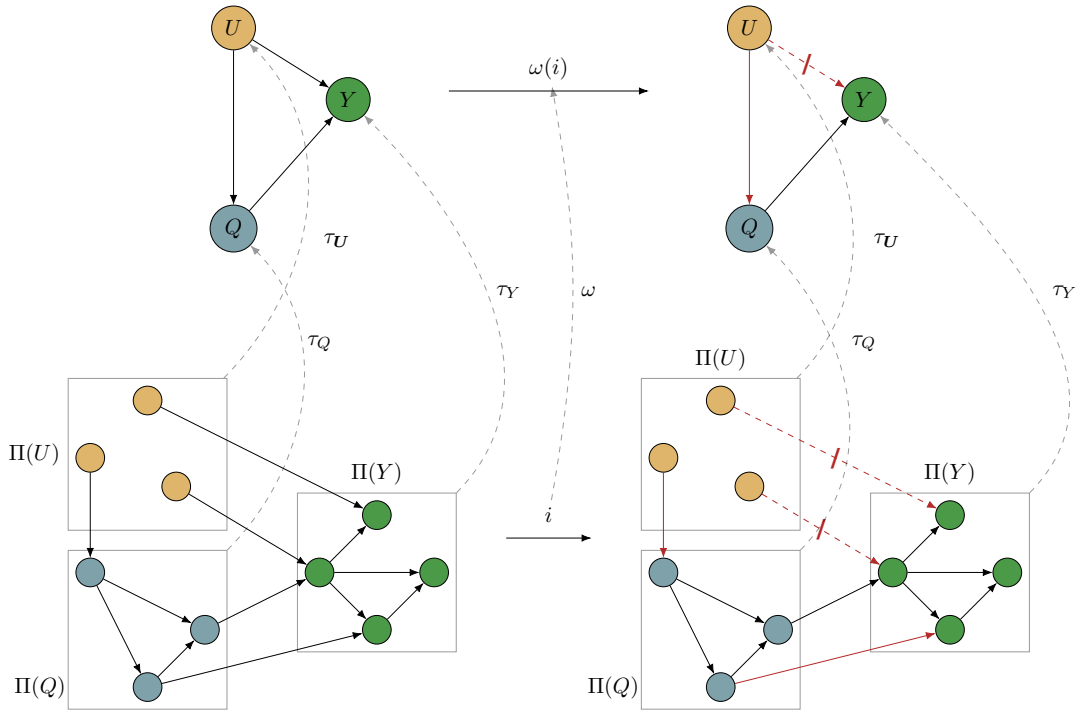

Figure 1: A visual depiction of constructive soft abstraction. The alignment between the high and low-level models is given by a map Π from high-level variables to sets of low-level variables and functions τ mapping exogenous and endogenous settings. The surjective partial function ω maps a low-level intervention i to a high-level intervention $\omega(i)$. Notably, the function ω is uniquely induced by the functions τ, and we provide its explicit form. Interventions may be hard (indicated by the red slash through the dashed red arrow) or soft (indicated by a solid red arrow with no slash). Observe that an intervention being performed on the connection between low-level clusters does not necessarily result in a high-level interventions.

2. Related Work

Rubenstein et al. (2017) introduce the notion of an "exact transformation" to understand when a probabilistic causal model can be transformed into another model in a causally consistent way. The core mathematical objects in an exact transformation are τ, a map between total settings of the causal models, and ω, a surjective and order-preserving map between hard interventions. Beckers and Halpern (2019) build on this work, introducing the notion of τ-abstraction, which requires τ to induce a specific function ω in order to rule out unintuitive aspects of exact transformation. Our work directly extends their definitions to handle soft interventions. Since our work focuses on redefining foundations of causal abstraction for soft interventions, we do not tackle the problem of measuring approximate abstraction (Beckers et al., 2019; Rischel and Weichwald, 2021). For an in depth survey of similar abstraction relations, we refer the reader to Zennaro (2022), which reviews various causal abstraction proposal for hard intervened models.

3. Structural Causal Models

Throughout the paper, we use notation $\boldsymbol{X}$ to denote a fixed set of variables, each $V \in \boldsymbol{X}$ with domain $\mathrm{Val}(V)$ of possible values. We refer to set of all possible subsets of variables as the power set $\mathcal{P}(\boldsymbol{X})$. The domain of any subset of variables $\boldsymbol{V} \subseteq \boldsymbol{X}$ is the Cartesian product of the members domains $\mathrm{Val}(\boldsymbol{V}) = \bigtimes_{V \in \boldsymbol{V}} \mathrm{Val}(V)$. We refer to the values $\boldsymbol{v} \in \mathrm{Val}(\boldsymbol{V})$ as *partial settings*. Similarly, we call $\boldsymbol{x} \in \mathrm{Val}(\boldsymbol{X})$ a *total setting*. Finally, given a total setting $\boldsymbol{x} \in \mathrm{Val}(\boldsymbol{X})$ and a set of variables $\boldsymbol{V} \subseteq \boldsymbol{X}$, we define the projection $\mathrm{Proj}(\boldsymbol{x}, \boldsymbol{V}) \in \mathrm{Val}(\boldsymbol{V})$ as the restriction of $\boldsymbol{x}$ to the variables $\boldsymbol{V}$.

Structural Causal Models (SCMs) express causal connections as functional relations between variables (Bongers et al., 2021). In particular, SCMs distinguish between endogenous variables, for which the causal mechanism is determined, and exogenous independent random variables.

Definition 1 (Structural Causal Model) *A Structural Causal Model $\mathcal{M}$ is a tuple $(\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$ such that*

1. *$\boldsymbol{X}$ is a set of endogenous variables with domain $\mathrm{Val}(X)$ for each $X \in \boldsymbol{X}$,*
2. *$\boldsymbol{E}$ is a set of exogenous variables with domain $\mathrm{Val}(E)$ for each $E \in \boldsymbol{E}$,*
3. *$\boldsymbol{F}$ is a set of functions with form $F_X \colon \mathrm{Val}(\boldsymbol{X}) \times \mathrm{Val}(\boldsymbol{E}) \to \mathrm{Val}(X)$ for each $X \in \boldsymbol{X}$,*
4. *$\mathbb{P}_E$ is a probability measure on $\mathrm{Val}(\boldsymbol{E})$.*

For each endogenous variable $X \in \boldsymbol{X}$, the structural equation F_X depends only on a subset of variables. We refer to this subset as the parents $\mathrm{Pa}(X) \subseteq \boldsymbol{X} \cup \boldsymbol{E}$ of the variable X. Endogenous settings are fully determined by exogenous settings. Therefore, each model can be considered as a function $\mathcal{M} \colon \mathrm{Val}(\boldsymbol{E}) \to \mathrm{Val}(\boldsymbol{X})$ from exogenous to endogenous values.

An intervention describes an operation on an SCM that produces a possibly different causal model by modifying its structural equations. We refer to interventions fixing structural equations to constant values as "hard" interventions (Pearl, 2009).

Definition 2 (Hard Intervention) *Given an SCM $\mathcal{M} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_E)$, a subset of endogenous variables $\boldsymbol{V} \subseteq \boldsymbol{X}$, and a partial setting $\boldsymbol{v} \in \mathrm{Val}(\boldsymbol{V})$, a hard intervention $i = (\boldsymbol{V} \leftarrow \boldsymbol{v})$ produces an SCM $\mathcal{M}_i = (\boldsymbol{X}, \boldsymbol{U}, \boldsymbol{F}^i, \mathbb{P}_E)$ such that*

$$F_X^i(\boldsymbol{x}, \boldsymbol{e}) = \begin{cases} \mathrm{Proj}(\boldsymbol{v}, X) & X \in \boldsymbol{V} \\ F_X(\boldsymbol{x}, \boldsymbol{e}) & \textit{otherwise} \end{cases} \tag{1}$$

for each endogenous variable $X \in \boldsymbol{X}$,

By constraining one or more endogenous variables, a hard intervention reduces the set of values possibly assumed by an SCM. Following Beckers and Halpern (2019), given an intervention on a model, we refer to the resulting subset of assumable values as its restriction set.

Definition 3 (Restriction Set) *The restriction $\mathrm{Rst}(\mathcal{M}_i)$ of a hard intervention $i = (\boldsymbol{V} \leftarrow \boldsymbol{v})$ on a model $\mathcal{M}$ is the subset of total settings on $\boldsymbol{X} \supseteq \boldsymbol{V}$ matching the partial setting $\boldsymbol{v}$. Formally,*

$$\mathrm{Rst}(\mathcal{M}_i) = \{\boldsymbol{x} \in \mathrm{Val}(\boldsymbol{X}) \mid \boldsymbol{v} = \mathrm{Proj}(\boldsymbol{x}, \boldsymbol{V})\}. \tag{2}$$

4. Causal Abstraction with Hard Interventions

Intuitively, Beckers and Halpern (2019) define an abstraction between two causal models whenever a function on their variables settings induces a function between interventions. The intervention map must be consistent: intervening on the higher-level abstraction should be equivalent to abstracting the intervened lower-level model.

Definition 4 (Abstraction) *Let $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$ and $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ be two SCMs, $\mathcal{I}$ and $\mathcal{J}$ be their respective sets of admissible hard interventions, $\tau_{\boldsymbol{Y}} \colon \mathrm{Val}(\boldsymbol{X}) \to \mathrm{Val}(\boldsymbol{Y})$ be a surjective function between endogenous settings, and $\tau_{\boldsymbol{U}} \colon \mathrm{Val}(\boldsymbol{E}) \to \mathrm{Val}(\boldsymbol{U})$ be a surjective function between exogenous settings. We can use $\tau_{\boldsymbol{Y}}$ to induce $\omega \colon \mathcal{I} \to \mathcal{J}$, a partial function from low-level hard interventions to high-level hard interventions. Specifically, ω is defined on the low-level intervention $i \in \mathcal{I}$ whenever there exists an high-level intervention $\omega(i) \in \mathcal{J}$ such that*

$$\mathrm{Rst}(\mathcal{H}_{\omega(i)}) = \{\tau_{\mathbf{Y}}(\mathbf{x}) \mid \mathbf{x} \in \mathrm{Rst}(\mathcal{L}_i)\}. \tag{3}$$

The model $\mathcal{H}$ is a τ-abstraction of $\mathcal{L}$ whenever ω is surjective and, for any intervention $i \in \mathcal{I}$ and exogenous setting $\boldsymbol{e} \in \boldsymbol{E}$ at the lower level, it holds

$$\tau_{\boldsymbol{Y}}(\mathcal{L}_i(\boldsymbol{e})) = \mathcal{H}_{\omega(i)}(\tau_{\boldsymbol{U}}(\boldsymbol{e})). \tag{4}$$

Given Definition 4, whenever the function ω between interventions exists, it is unique. Furthermore, by ensuring consistency on all exogenous settings, abstraction entails that the high-level model is an exact transformation (Rubenstein et al., 2017) between SCMs for every possible exogenous distribution. Therefore, whenever a model abstracts another, the intervention mapping ω has a fixed point in the empty intervention and preserves the following intervention ordering.

Definition 5 (Hard Interventions Ordering) *A hard intervention $i_1 = (\boldsymbol{V} \leftarrow \boldsymbol{v})$ on an SCM $\mathcal{M}$ precedes another intervention $i_2 = (\boldsymbol{W} \leftarrow \boldsymbol{w})$ whenever the former sets to the same values a subset of the variables intervened by the latter. Formally,*

$$i_1 \sqsubseteq i_2 \iff \boldsymbol{V} \subseteq \boldsymbol{W} \wedge \boldsymbol{v} = \mathrm{Proj}(\boldsymbol{w}, \boldsymbol{V}). \tag{5}$$

5. Causal Abstraction with Soft Interventions

In this section, we extend τ-abstraction to relate causal models according to their response to non-constant interventions on their endogenous variables. Firstly, we generalize the necessary definitions of restriction set and partial ordering between hard interventions (Subsection 5.1). Then, we define low soft abstraction to handle soft interventions on the low-level model (Subsection 5.2). In this context, we introduce a running example showing how our definition correctly characterizes low-level soft abstracted models. Then, we prove that this preliminary definition is ambiguous for soft interventions on the higher-level model. Therefore, we modify our example to display an ambiguous scenario. Finally, we present our definition of soft abstraction, which correctly disambiguates higher-level soft interventions (Subsection 5.3). Consequently, we showcase how our contribution correctly solves the critical scenario in our running example.

5.1. Soft Interventions

Existing definitions of abstraction explicitly build on "hard" interventions that fix the structural equation of each intervened variable to a constant. A *soft* intervention is a generalization of an hard intervention that instead replaces structural equations with new, possibly non-constant functions.

Definition 6 (Soft Intervention) *Given an SCM $\mathcal{M} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$, a subset of endogenous variables $\boldsymbol{V} \subseteq \boldsymbol{X}$, and a set of functions $\boldsymbol{f}$, a soft intervention $i = (\boldsymbol{V} \leftarrow \boldsymbol{f})$ produces an SCM $\mathcal{M}_i = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}^i, \mathbb{P}_{\boldsymbol{E}})$ such that*

$$F_X^i(\boldsymbol{x}, \boldsymbol{e}) = \begin{cases} f_X(\boldsymbol{x}, \boldsymbol{e}) & X \in \boldsymbol{V} \\ F_X(\boldsymbol{x}, \boldsymbol{e}) & \textit{otherwise} \end{cases} \tag{6}$$

for each endogenous variable $X \in \boldsymbol{X}$.

For our purposes, we will impose two constraints on soft interventions. First, for each variable $X \in \boldsymbol{V}$ the function f_X must share domain and codomain with the non-intervened structural equation F_X. In this way, we enforce that a soft intervention does not alter the "type" of a variable. Second, we require intervened structural equations to depend at most on the same parents $\mathrm{Pa}(X)$ from the non-intervened model, meaning no new causal dependencies are formed between variables. Given this formalization, hard interventions are a special case of soft interventions where structural equations are replaced by constant functions.

To define causal abstraction between soft-intervened models, we must first extend the intervention map ω to be a partial function between soft interventions. Abstraction uniquely induces such function given the restrictions of hard interventions on the lower- and higher-level models. To this end, we propose the following generalization of restriction set for soft interventions on an SCM.

Definition 7 (Soft Restriction Set) *The soft restriction $\mathrm{SoftRst}(\mathcal{M}_i)$ of a soft intervention $i = (\boldsymbol{V} \leftarrow \boldsymbol{f})$ on a model $\mathcal{M}$ is the subset of total settings on $\boldsymbol{X} \supseteq \boldsymbol{V}$ matching any partial setting from the image of $\boldsymbol{f}$. Formally,*

$$\mathrm{SoftRst}(\mathcal{M}_i) = \{\boldsymbol{x} \in \mathrm{Val}(\boldsymbol{X}) \mid \mathrm{Proj}(\boldsymbol{x}, V) \in \mathrm{Image}(\boldsymbol{f})\} \tag{7}$$

Previous work introduced partial ordering between hard interventions according to constants assigned to intervened variables (Rubenstein et al., 2017). Furthermore, preserving intervention ordering constitutes a desirable property of causal abstraction. Therefore, we propose the following partial ordering on soft interventions.

Definition 8 (Soft Interventions Ordering) *Given two soft interventions i_1 and i_2 on an SCM $\mathcal{M}$, the intervention i_1 precedes i_2 whenever the soft restriction of the former contains that of the latter. Formally,*

$$i_1 \preceq i_2 \iff \mathrm{SoftRst}(\mathcal{M}_{i_1}) \supseteq \mathrm{SoftRst}(\mathcal{M}_{i_2}). \tag{8}$$

Our definition of restriction for soft interventions is a strict generalization of the corresponding definition on hard interventions. Our partial ordering on soft interventions is also a strict generalization of the ordering on hard interventions. See Appendix A.1 for proofs.

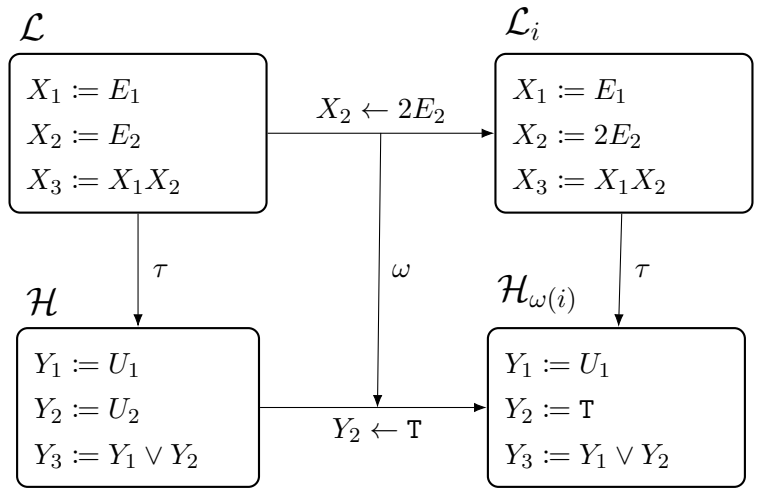

Figure 2: The low-level model $\mathcal{L}$ multiplies two integers. The high level model $\mathcal{H}$ computes whether one of two Boolean variables are true. The model $\mathcal{H}$ abstracts the model $\mathcal{L}$ with a map τ that computes the parity of each integer at the low-level, where X_n determines Y_n. This is because when two integers are multiplied, the product is even if and only if at least one of the integers is even. The soft intervention $i = (X_2 \leftarrow 2E_2)$ doubles the value of a low-level variable X_2 making it always even. Therefore, it is mapped to a constant hard intervention $\omega(i) = (Y_2 \leftarrow \mathtt{T})$ that fixes the high-level variable Y_2 to be always true.

5.2. Low-Level Soft Interventions

The definition of τ-abstraction induces a unique transformation ω between interventions whenever the restrictions of any pair of lower- and higher-level interventions are related by τ. Given our generalization, we can use soft restriction in place of strict restriction for a definition of abstraction immediately supporting low-level soft interventions.

Definition 9 (Low Soft Abstraction) *Let $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$ and $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ be two SCMs, $\mathcal{I}$ and $\mathcal{J}$ be their respective sets of admissible interventions, $\tau_{\boldsymbol{Y}} \colon \mathrm{Val}(\boldsymbol{X}) \to \mathrm{Val}(\boldsymbol{Y})$ be a surjective function between endogenous settings, and $\tau_{\boldsymbol{U}} \colon \mathrm{Val}(\boldsymbol{E}) \to \mathrm{Val}(\boldsymbol{U})$ be a surjective function between exogenous settings. The model $\mathcal{H}$ is a low soft τ-abstraction of $\mathcal{L}$ whenever it exists a surjective function $\omega \colon \mathcal{I} \to \mathcal{J}$ such that, for any intervention $i \in \mathcal{I}$ it holds*

$$\mathrm{SoftRst}(\mathcal{H}_{\omega(i)}) = \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_i)\} \tag{9}$$

and for every lower-level exogenous setting $\boldsymbol{e} \in \boldsymbol{E}$, it holds

$$\tau_{\boldsymbol{Y}}(\mathcal{L}_i(\boldsymbol{e})) = \mathcal{H}_{\omega(i)}(\tau_{\boldsymbol{U}}(\boldsymbol{e})). \tag{10}$$

As proved in Appendix A.2, by assuming that the high-level model admits only hard interventions, the induced mapping ω is uniquely defined, order-preserving, and has a fixed point in the empty intervention. Furthermore, whenever admissible interventions on both models are hard, low

soft τ-abstraction reduces to τ-abstraction. As the following example shows, Definition 9 effectively represents abstraction whenever a low-level soft intervention results in an hard intervention on the abstract model.

Example 1 *Let $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$ and $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ be two SCMs with structural equations as reported in Figure 2. At the lower-level, every endogenous or exogenous variable $X \in \boldsymbol{X} \cup \boldsymbol{E}$ is integer with domain $\mathrm{Val}(X) = \mathbb{Z}$. The high-level model $\mathcal{H}$ consists instead of Boolean variables $Y \in \boldsymbol{Y} \cup \boldsymbol{U}$ with domain $\mathrm{Val}(Y) = \mathbb{B}$. Given a function τ composed by element-wise mapping a function $\mathrm{Even}\colon \mathbb{Z} \to \mathbb{B}$ testing the parity of each integer, $\mathcal{H}$ is a τ-abstraction of $\mathcal{L}$ for every hard intervention at the lower-level. Furthermore, any soft intervention at the lower-level that always produces either even or odd integers also results in an abstract hard intervention. For instance, let us take the soft intervention $i = (X_2 \leftarrow 2E_2)$: its soft restriction $\mathrm{SoftRst}(\mathcal{L}_i)$ is the set of integer triples $\{(x_1, x_2, x_3) \mid \mathrm{Even}(x_2) = \mathtt{T}\}$. Consequently, the result of applying τ to the soft restriction of i is the set of Boolean triples $\{y_1, y_2, y_3 \mid y_2 = \mathtt{T}\}$. The set coincides exclusively with the restriction of the high-level hard intervention $j = (Y_2 \leftarrow \mathtt{T})$. Furthermore, the consistency requirement holds as in*

$$\begin{aligned}
\tau_{\boldsymbol{Y}}(\mathcal{L}_i([e_1, e_2])) &= \tau_Y([e_1, 2e_2, e_1e_2]) \\
&= [\mathrm{Even}(e_1), \mathtt{T}, \mathrm{Even}(e_1e_2)] \\
&= [\mathrm{Even}(e_1), \mathtt{T}, \mathrm{Even}(e_1) \vee \mathrm{Even}(e_2)] \\
&= \mathcal{H}_j(\tau_{\boldsymbol{U}}([e_1, e_2])).
\end{aligned} \tag{11}$$

Therefore, $\mathcal{H}$ low soft abstracts $\mathcal{L}$ and ω uniquely maps the low level intervention i to j.

Crucially, the intervention map ω is only guaranteed to be unique whenever the set of admissible interventions at the higher level contains exclusively hard interventions. The definition of abstraction provided by Beckers and Halpern (2019) ensures consistency for all exogenous settings. Nonetheless, whenever structural equations are non-surjective, consistency is not enforced on all endogenous settings. Therefore, an intervention on the low level model might be mapped to multiple distinct abstract soft interventions that coincide for any exogenous-induced endogenous setting. Consequently, as proved in the following theorem, there exist different maps ω for each possible choice of high-level intervention.

Theorem 10 (Ambiguity in Low Soft Abstraction) *Let $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ be an SCM with admissible interventions $\mathcal{J}$. Given a variable $V \in \boldsymbol{Y}$, if $\mathcal{J}$ contains two distinct interventions $j = (V \leftarrow g)$ and $j' = (V \leftarrow g')$ such that*

1. $\forall \boldsymbol{u} \in \mathrm{Val}(\boldsymbol{U})\colon \mathrm{Proj}(\mathcal{H}_j(\boldsymbol{u}), V) = \mathrm{Proj}(\mathcal{H}_{j'}(\boldsymbol{u}), V)$,
2. $\mathrm{Image}(g) = \mathrm{Image}(g')$,

then, for any causal model $\mathcal{L}$ with admissible interventions $\mathcal{I}$, whenever $\mathcal{H}$ low soft τ-abstracts $\mathcal{L}$, the corresponding function $\omega\colon \mathcal{I} \to \mathcal{J}$ is not uniquely defined.

Proof We report the proof in Appendix A.3. ■

In the following, we alter the previous example to report a scenario where non-surjective structural equations lead to a possibly non-unique intervention map ω.

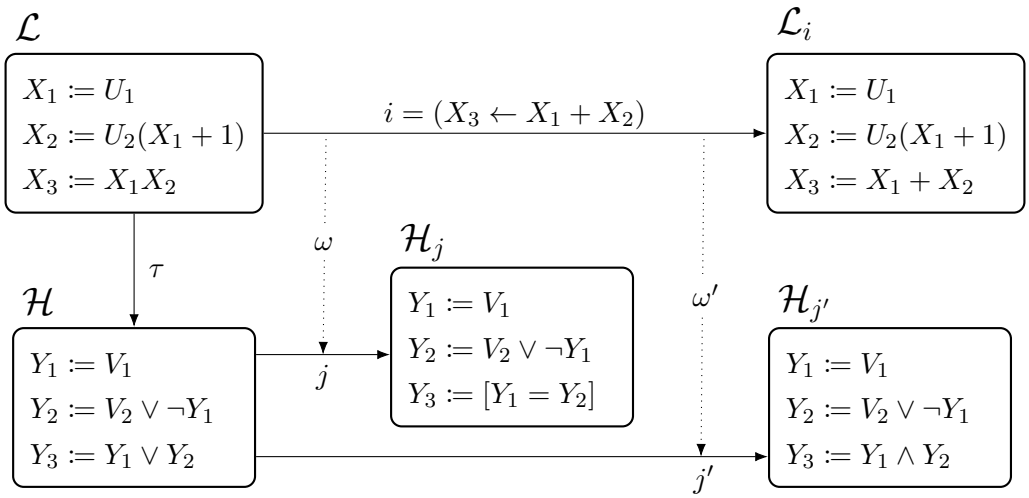

Figure 3: The low-level model $\mathcal{L}$ performs arithmetic operations on two integers. The high level model $\mathcal{H}$ computes a Boolean proposition. The model $\mathcal{H}$ abstracts the model $\mathcal{L}$ with a map τ that computes the parity of each integer at the low-level, where X_n determines Y_n. Given the intervention $i = (X_3 \leftarrow X_1 + X_2)$, there exist two distinct intervention maps satisfying low soft abstraction: $\omega(i) = (Y_3 \leftarrow [Y_1 = Y_2])$ and $\omega'(i) = (Y_3 \leftarrow Y_1 \wedge Y_2)$, where the former tests equality. Intuitively, this occurs since the two interventions differ only when Y_1 and Y_2 are both false, but this condition is never reached for any exogenous setting.

Example 2 *Let $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$ and $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ be two SCMs with structural equations as reported in Figure 3. As in previous Example 1, Boolean-valued model $\mathcal{H}$ τ-abstracts integer-valued model $\mathcal{L}$ given a function τ element-wise testing the parity of each integer. Differently from before, the structural equations of X_1 and X_2 are jointly not surjective. In fact, for any exogenous setting, X_1 and X_2 will never be both odd. Similarly, at the higher-level, variables Y_1 and Y_2 will never be both false. We define a low-level soft intervention $i = (X_3 \leftarrow X_1 + X_2)$ and two high-level soft interventions $j = (Y_3 \leftarrow [Y_1 = Y_2])$, $j' = (Y_3 \leftarrow Y_1 \wedge Y_2)$. Exploiting the unrealized endogenous values, we can show that exist a function ω mapping i to j and a function ω' mapping i to j', both satisfying the requirements for low soft abstraction.*

The function for intervention i outputs even ad odd numbers and the functions for j and j' output both true and false values, so we know that

$$\mathrm{SoftRst}(\mathcal{H}_{j'}) = \mathrm{SoftRst}(\mathcal{H}_j) = \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_i)\}.$$

Furthermore, for any low-level exogenous setting $\boldsymbol{e} \in \boldsymbol{E}$, it holds that

$$\tau_{\boldsymbol{Y}}(\mathcal{L}_i(\boldsymbol{e})) = \mathcal{M}_j(\tau_{\boldsymbol{U}}(\boldsymbol{e}))$$
$$\tau_{\boldsymbol{Y}}(\mathcal{L}_i(\boldsymbol{e})) = \mathcal{M}_{j'}(\tau_{\boldsymbol{U}}(\boldsymbol{e})).$$

Consequently, given a function ω witnessing that $\mathcal{H}$ low soft τ-abstracts $\mathcal{L}$, we can construct a distinct map ω' that differs on input i and also witnesses the same abstraction relationship.

5.3. High-Level and Low-Level Soft Interventions

Intuitively, ambiguity in low soft abstraction arises whenever there exist two distinct soft interventions on a variable whose parents have non-surjective structural equations. Therefore, two distinct functions could be assigned to the same variable if they agree at least on the image of the structural equations of their parents. We address this problem by requiring consistency not only for all exogenous settings, but also for every endogenous settings.

Definition 11 (Soft Abstraction) *Let $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$ and $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ be two SCMs, $\mathcal{I}$ and $\mathcal{J}$ be their respective sets of admissible interventions, $\tau_{\boldsymbol{Y}} \colon \mathrm{Val}(\boldsymbol{X}) \to \mathrm{Val}(\boldsymbol{Y})$ be a surjective function between endogenous settings, and $\tau_{\boldsymbol{U}} \colon \mathrm{Val}(\boldsymbol{E}) \to \mathrm{Val}(\boldsymbol{U})$ be a surjective function between exogenous settings. The model $\mathcal{H}$ is a soft τ-abstraction of $\mathcal{L}$ whenever it exists a surjective function $\omega \colon \mathcal{I} \to \mathcal{J}$ such that, for any intervention $i \in \mathcal{I}$ it holds*

$$\mathrm{SoftRst}(\mathcal{H}_{\omega(i)}) = \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_i)\} \tag{12}$$

and for every lower-level exogenous $\boldsymbol{e} \in \boldsymbol{E}$ and endogenous $\boldsymbol{x} \in \boldsymbol{X}$ setting, it holds

$$\tau_{\boldsymbol{Y}}(\boldsymbol{F}^i(\boldsymbol{x}, \boldsymbol{e})) = \boldsymbol{G}^{\omega(i)}(\tau_{\boldsymbol{Y}}(\boldsymbol{x}), \tau_{\boldsymbol{U}}(\boldsymbol{e})). \tag{13}$$

Given this definition, the intervention mapping ω is unique for every combination of hard and soft interventions. Notably, our definition specializes to low soft τ-abstraction when all high-level interventions are hard. Similarly, soft τ-abstraction implies τ-abstraction, and vice versa, when all admitted interventions are hard interventions. We report in Appendix A.4 the proofs on uniqueness, fixed point, and order-preserving properties of the intervention mapping ω in soft abstraction. Further, we continue previous Example 2 to highlight how soft abstraction disambiguates high-level soft interventions.

Example 3 *Recall the previous Example 2, where there were two intervention maps ω and ω' that, given the two SCMs in Figure 3, witnessed $\mathcal{H}$ being an abstraction of $\mathcal{L}$. Given the exogenous setting $\boldsymbol{e} = (1, 1)$ and the endogenous setting $\boldsymbol{x} = (1, 1, 1)$, it is immediate that*

$$\begin{aligned} \tau_{\boldsymbol{Y}}(\mathbf{F}^i(\boldsymbol{x}, \boldsymbol{e})) &= [\mathtt{F}, \mathtt{T}, \mathtt{T}] \\ \mathbf{G}^{\mathbf{j}}(\tau_{\mathbf{Y}}(\mathbf{x}, \tau_{\mathbf{U}}(\mathbf{e}))) &= [\mathtt{F}, \mathtt{T}, \mathtt{T}] \\ \mathbf{G}^{\mathbf{j}'}(\tau_{\mathbf{Y}}(\mathbf{x}, \tau_{\mathbf{U}}(\mathbf{e}))) &= [\mathtt{F}, \mathtt{T}, \mathtt{F}]. \end{aligned} \tag{14}$$

Therefore, mapping the intervention i to j' would break consistency on the provided setting. Thus, we no longer have an ambiguous intervention map ω.

6. Constructive Soft Abstraction

Our definition of Soft Abstraction relates two distinct SCMs at different levels of detail according to their response to possibly soft interventions. In this way, we can determine abstraction whenever an intervention at the lower level produces a non-constant effect at the higher level, vice versa, or both. One critical result is the uniqueness of the function mapping interventions between SCMs. Therefore, whenever a model abstracts another, each lower-level manipulation corresponds to a unique abstract intervention. In this section, we build on the uniqueness of the intervention mapping to derive its closed and explicit form.

To this end, we require grouping endogenous and exogenous variables on the lower-level model. Therefore, we consider the existence of an alignment function mapping high-level variables into subsets of low-level variables.

Definition 12 (Alignment) *Given two SCMs $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ and $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$ such that $\mathcal{H}$ soft τ-abstracts $\mathcal{L}$, we define an alignment to be a map $\Pi\colon \boldsymbol{Y} \cup \boldsymbol{U} \to \mathcal{P}(\boldsymbol{X} \cup \boldsymbol{E})$ that satisfy the following: (1) There exist a set of functions $\{\tau_Y\}_{Y \in \boldsymbol{Y}}$ such that, for each endogenous high-level variable $Y \in \boldsymbol{Y}$ and low-level endogenous setting $\boldsymbol{x} \in \boldsymbol{X}$, it holds $\mathrm{Proj}(\tau_{\boldsymbol{Y}}(\boldsymbol{x}), Y) = \tau_Y(\mathrm{Proj}(\boldsymbol{x}, \Pi(Y)))$, where $\Pi(Y) \subseteq \boldsymbol{X}$ and $\tau_Y\colon \mathrm{Val}(\Pi(Y)) \to \mathrm{Val}(Y)$; (2) There exist a set of functions $\{\tau_U\}_{U \in \boldsymbol{U}}$ such that, for each exogenous high-level variable $U \in \boldsymbol{U}$ and low-level exogenous setting $\boldsymbol{e} \in \boldsymbol{E}$ it holds $\mathrm{Proj}(\tau_{\boldsymbol{U}}(\boldsymbol{e}), U) = \tau_U(\mathrm{Proj}(\boldsymbol{e}, \Pi(U)))$, where $\Pi(U) \subseteq \boldsymbol{E}$ and $\tau_U\colon \mathrm{Val}(\Pi(U)) \to \mathrm{Val}(U)$.*

Assuming the existence of a map from high-level to low-level variables is coherent with previous definitions of constructive abstraction (Beckers and Halpern, 2019) and model transformation (Rischel and Weichwald, 2021) for hard-intervened causal models. In particular, constructive abstraction further specializes causal abstraction to address the problem of clustering micro-variables into distinct causal variables at the higher-level. Given our definition of Soft Abstraction, a similar extension is immediate in our framework. Therefore, we define Constructive Soft Abstraction whenever the abstraction partitions lower-level variables.

Definition 13 (Constructive Soft Abstraction) *An SCM $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ is a constructive soft τ-abstraction of another SCM $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$ whenever $\mathcal{H}$ τ-abstracts $\mathcal{L}$ and there exists an alignment Π partitioning a subset of low-level variables.*

Since constructive soft abstraction extends soft abstraction, the function ω between interventions is unique. Therefore, we seek an explicit form of ω returning, for an intervention $(\Pi(Y) \leftarrow \boldsymbol{f})$ on the partition of an high-level variable Y, the corresponding higher-level intervention $(Y \leftarrow g)$. With the following theorem, we prove that the target function g_Y necessarily consists of the composition of three operations: (i) mapping the inputs of the function on the partitions of the parents $\Pi(\mathrm{Pa}(Y))$; (ii) computing the value of the partition $\Pi(Y)$ on the intervened model $\mathcal{L}_i$; and (iii) finally mapping back the partition to the higher-level variable Y (Figure 1).

Theorem 14 (Explicit Intervention Map) *Let $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$ and $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ be two SCMs and $\mathcal{I}$ and $\mathcal{J}$ their corresponding sets of admissible interventions. Whenever $\mathcal{H}$ soft τ-abstracts $\mathcal{L}$ with alignment Π, ω maps the low-level intervention $i = (\boldsymbol{V} \leftarrow \boldsymbol{f})$ to an intervention $j = (\boldsymbol{W} \leftarrow \boldsymbol{g})$ on the high-level variables $\boldsymbol{W} \subseteq \{Y \mid \boldsymbol{V} \cap \Pi(Y) \neq \emptyset\}$ such that*

$$g_Y(\boldsymbol{y}, \boldsymbol{u}) = \tau_Y(F^i_{\Pi(Y)}(\tau^{-1}_{\boldsymbol{Y}}(\boldsymbol{y}), \tau^{-1}_{\boldsymbol{U}}(\boldsymbol{u}))), \tag{15}$$

for any variable $Y \in \boldsymbol{W}$ and any partial inverse τ^{-1}, where g_Y depends on its parents $\mathrm{Pa}(Y)$ only.

Proof We report the proof in Appendix A.5. ■

Remarkably, Theorem 14 does not require the alignment Π to partition low-level variables. Consequently, it applies to constructive soft abstraction as a special case. Furthermore, whenever both sets of admissible interventions are hard, the explicit form is a necessary condition for constructive τ-abstraction. In the following example, we showcase how the explicit definition of the intervention map ω can be applied to determine high-level interventions.

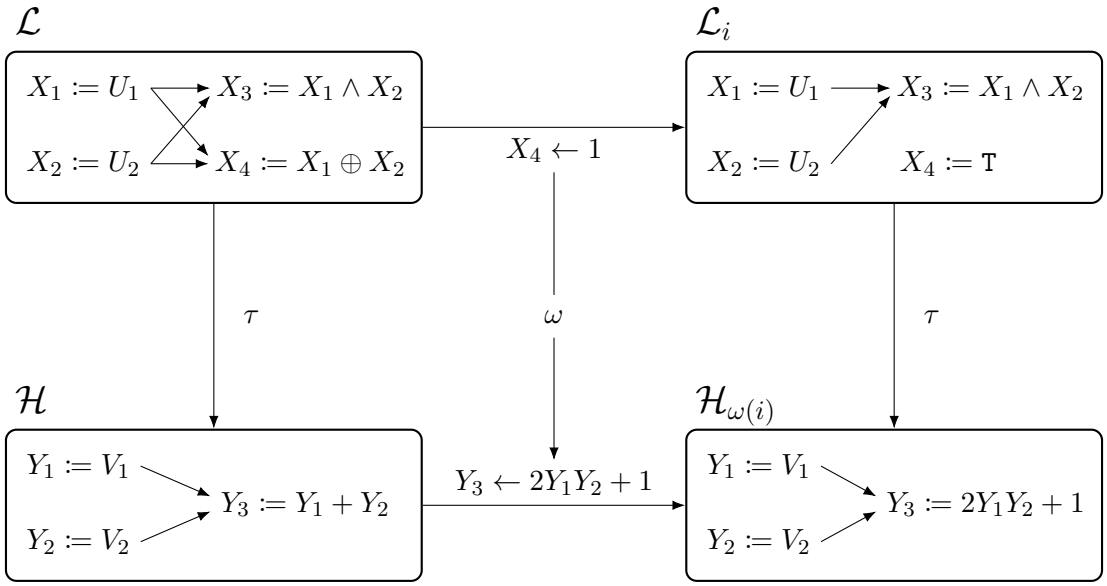

Figure 4: The SCM $\mathcal{L}$ models a binary adder of two exogenous bits U_1, U_2, where $\oplus$ is the Boolean XOR operator. The model $\mathcal{H}$ adds two integers and stores the sum in a single integer Y_3. Performing an hard intervention on $\mathcal{L}$, for instance by fixing the least significant bit $X_4 \leftarrow \mathtt{T}$, results in a soft intervention at the higher-level $Y_3 \leftarrow 2Y_1Y_2 + 1$.

Example 4 *Let $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$ and $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ be two SCMs with structural equations as reported in Figure 4. At the low-level, every endogenous and exogenous variable $X \in \boldsymbol{X} \cup \boldsymbol{E}$ has Boolean domain $\mathrm{Val}(X) = \{\mathtt{F}, \mathtt{T}\}$. At the high level, $\mathrm{Val}(Y_3) = \{0, 1, 2, 3\}$, while every other exogenous or endogenous variable $Y \in \boldsymbol{Y} \cup \boldsymbol{U} \setminus \{Y_3\}$ has domain $\mathrm{Val}(Y) = \{0, 1\}$. We consider the alignment Π where $\Pi(Y_3) = \{X_3, X_4\}$, while $\Pi(Y_n) = \{X_n\}$ for every other high-level exogenous or endogenous variable Y_n. Further, we define the abstraction as $\tau_{Y_3}(\boldsymbol{x}) = 2\varphi(x_3) + \varphi(x_4)$, and $\tau_{Y_n} = \varphi$ for the remaining variables Y_n, where φ maps true values to the integer 1, and false values to 0. Given the low-level intervention $i = (X_4 \leftarrow \mathtt{T})$, the corresponding high-level intervention is $\omega(i) = (Y_3 \leftarrow g)$, where*

$$
\begin{aligned}
g(\boldsymbol{y}, \boldsymbol{u}) &= \tau_{Y_3}(F^i_{\Pi(Y_3)}(\tau^{-1}_{\boldsymbol{Y}}(\boldsymbol{y}), \tau^{-1}_{\boldsymbol{U}}(\boldsymbol{u}))) \\
&= \tau_{Y_3}(F^i_{\{X_3, X_4\}}((\varphi^{-1}(y_1), \varphi^{-1}(y_2), x_3, x_4), \boldsymbol{e})) \\
&= \tau_{Y_3}(\varphi^{-1}(y_1) \wedge \varphi^{-1}(y_2), \mathtt{T}) \\
&= 2\varphi(\varphi^{-1}(y_1) \wedge \varphi^{-1}(y_2)) + \varphi(\mathtt{T}) \\
&= 2y_1y_2 + 1
\end{aligned}
\tag{16}
$$

Since g depends only on its parents, the explicit form works for any choice of the remaining endogenous $\{X_3, X_4\}$ or exogenous $\boldsymbol{E}$ lower-level variables.

7. Conclusion

Hard interventions constitute one of the core concepts of causality. Existing theories of abstraction build on the assumption that a causal model abstracts another whenever there is a correspondence between their sets of admissible hard interventions. We generalize this theory to soft interventions to support non-constant change at different levels of abstraction. To this end, we prove how generalizing constructs based on hard interventions in the definition of τ-abstraction from Beckers and Halpern (2019) does not suffice. Therefore, we propose to enforce causal consistency for all exogenous *and* endogenous settings. Together with our generalization effort, this strengthened requirement results in our novel definition of soft abstraction. As proved and exemplified, soft abstraction induces a unique function ω mapping low-level interventions to high-level interventions in the soft-intervened scenario. Finally, we report that our constructive definition of soft abstraction implies a unique explicit form for the intervention map ω. Since soft abstraction correctly reduces to τ-abstraction in the hard-intervened scenario, this result effectively extends previous formulations that exclusively proved the uniqueness of ω. By demonstrating the explicit form as a necessary condition, our contribution opens up practical applications for assessing abstraction without requiring an evaluation of each exogenous or endogenous configuration.

References

Elias Bareinboim, Juan Correa, Duligur Ibeling, and Thomas Icard. On Pearl's hierarchy and the foundations of causal inference. In Hector Geffner, Rina Dechter, and Joseph Y. Halpern, editors, *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 509–556. ACM Books, 2022.

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 2678–2685, 2019.

Sander Beckers, Frederick Eberhardt, and Joseph Y. Halpern. Approximate causal abstractions. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, 2019.

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.

Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. In Marina Meila and Tom Heskes, editors, *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, UAI 2015, July 12-16, 2015, Amsterdam, The Netherlands*, pages 181–190. AUAI Press, 2015. URL http://auai.org/uai2015/proceedings/papers/109.pdf.

Krzysztof Chalupka, Tobias Bischoff, Frederick Eberhardt, and Pietro Perona. Unsupervised discovery of el nino using causal feature learning on microlevel climate data. In Alexander T. Ihler and Dominik Janzing, editors, *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence, UAI 2016, June 25-29, 2016, New York City, NY, USA*. AUAI Press, 2016. URL http://auai.org/uai2016/proceedings/papers/11.pdf.

Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44:137–164, 2017.

Julien Dubois, Frederick Eberhardt, Lynn K. Paul, and Ralph Adolphs. Personality beyond taxonomy. *Nature human behaviour*, 4 11:1110–1117, 2020a.

Julien Dubois, Hiroyuki Oya, Julian Michael Tyszka, Matthew A. Howard, Frederick Eberhardt, and Ralph Adolphs. Causal mapping of emotion networks in the human brain: Framework and initial findings. *Neuropsychologia*, 145, 2020b.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of science*, 74(5):981–995, 2007.

Atticus Geiger, Kyle Richardson, and Christopher Potts. Neural natural language inference models partially embed theories of lexical entailment and negation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 163–173, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.16. URL https://www.aclweb.org/anthology/2020.blackboxnlp-1.16.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

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

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

Eigil F Rischel and Sebastian Weichwald. Compositional abstraction error and a category of causal models. In *Uncertainty in Artificial Intelligence*, pages 1013–1023. PMLR, 2021.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. *arXiv preprint arXiv:1707.00819*, 2017.

James Woodward. The problem of variable choice, August 2015. URL http://philsci-archive.pitt.edu/11621/. Forthcoming in Synthese.

Fabio Massimo Zennaro. Abstraction between structural causal models: A review of definitions and properties. *arXiv preprint arXiv:2207.08603*, 2022.

Appendix A. Proofs

A.1. Soft Generalization of Hard Constructs

Given an SCM $\mathcal{M}$, for any hard intervention i on the model, it holds that

1. $\mathrm{SoftRst}(M_i) = \mathrm{Rst}(M_i)$,

   Proof Given a subset $\boldsymbol{V} \subseteq \boldsymbol{X}$ of endogenous variables, a hard intervention $i = (\boldsymbol{V} \leftarrow \boldsymbol{v})$ constraints each intervened variable X to a constant value v. Equivalently, we can consider the same intervention as a function $\boldsymbol{f}\colon \mathrm{Val}(\boldsymbol{X}) \times \mathrm{Val}(\boldsymbol{E}) \to \mathrm{Val}(\boldsymbol{V})$ whose output is the constant $\boldsymbol{v}$. Therefore, $\mathrm{Image}(\boldsymbol{f}) = \{v\}$ and, consequently,

$$\begin{aligned} \mathrm{SoftRst}(\mathcal{M}_i) &= \{\boldsymbol{x} \in \mathrm{Val}(\boldsymbol{X}) \mid \mathrm{Proj}(\boldsymbol{x}, V) \in \mathrm{Image}(\boldsymbol{f})\} \\ &= \{\boldsymbol{x} \in \mathrm{Val}(\boldsymbol{X}) \mid \mathrm{Proj}(\boldsymbol{x}, V) = \boldsymbol{v}\} \\ &= \mathrm{Rst}(\mathcal{M}_i). \end{aligned} \tag{17}$$

   ■

2. For any hard intervention i', $i \sqsubseteq i' \iff i \preceq i'$.

   Proof Given the previous property, we prove that soft intervention ordering implies hard intervention ordering for any pair of hard interventions $i = (\boldsymbol{V} \leftarrow \boldsymbol{v})$ and $i' = (\boldsymbol{W} \leftarrow \boldsymbol{w})$. In fact,

$$\begin{aligned} i \preceq i' &\iff \mathrm{SoftRst}(\mathcal{M}_i) \supseteq \mathrm{SoftRst}(\mathcal{M}_{i'}) \\ &\iff \mathrm{Rst}(\mathcal{M}_i) \supseteq \mathrm{Rst}(\mathcal{M}_{i'}), \end{aligned} \tag{18}$$

   whenever i, i' are hard interventions. If the restriction of $\mathcal{M}_i$ contains that of $\mathcal{M}_{i'}$, then $\mathcal{M}_i$ allows more values for some variables. Equivalently, i' intervenes on a larger set of variables than i and sets them to the same constants, otherwise they would not have been in the restriction of $\mathcal{M}_i$. Consequently,

$$\begin{aligned} \mathrm{Rst}(\mathcal{M}_i) \supseteq \mathrm{Rst}(\mathcal{M}_{i'}) &\iff \boldsymbol{V} \subseteq \boldsymbol{W} \wedge \boldsymbol{v} = \mathrm{Proj}(\boldsymbol{w}, \boldsymbol{V}) \\ &\iff i \sqsubseteq i'. \end{aligned} \tag{19}$$

   ■

A.2. Low Soft Abstraction

Definition 9: Let $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$ and $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ be two SCMs, $\mathcal{I}$ and $\mathcal{J}$ be their respective sets of admissible interventions, $\tau_{\boldsymbol{Y}}\colon \mathrm{Val}(\boldsymbol{X}) \to \mathrm{Val}(\boldsymbol{Y})$ be a surjective function between endogenous settings, and $\tau_{\boldsymbol{U}}\colon \mathrm{Val}(\boldsymbol{E}) \to \mathrm{Val}(\boldsymbol{U})$ be a surjective function between exogenous settings. The model $\mathcal{H}$ is a low soft τ-abstraction of $\mathcal{L}$ whenever it exists a surjective function $\omega\colon \mathcal{I} \to \mathcal{J}$ such that, for any intervention $i \in \mathcal{I}$ it holds

$$\mathrm{SoftRst}(\mathcal{H}_{\omega(i)}) = \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_i)\} \tag{20}$$

and for every lower-level exogenous setting $\boldsymbol{e} \in \boldsymbol{E}$, it holds

$$\tau_{\boldsymbol{Y}}(\mathcal{L}_i(\boldsymbol{e})) = \mathcal{H}_{\omega(i)}(\tau_{\boldsymbol{U}}(\boldsymbol{e})). \tag{21}$$

Given Definition 9 of Low Soft Abstraction, we prove that, whenever the set of admissible high-level intervention $\mathcal{J}$ contains only hard interventions and $\mathcal{H}$ low soft τ-abstracts $\mathcal{L}$, it holds that

1. ω is unique,

   Proof For any distinct surjective function $\omega' \neq \omega$, $\mathcal{H}$ does not τ-abstracts $\mathcal{L}$. Since ω, ω' are defined on the domain $\mathcal{I}$, they are distinct whenever there exist at least one intervention $i \in \mathcal{I}$ such that

   $$\omega(i) \neq \omega(i').$$

   Thus, there must exist two distinct high-level interventions $j, j' \in \mathcal{H}$ such that,

   $$\mathrm{SoftRst}(\mathcal{H}_j) = \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_i)\} \tag{22}$$
   $$\mathrm{SoftRst}(\mathcal{H}_{j'}) = \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_i)\} \tag{23}$$

   Consequently, it must hold that

   $$\mathrm{SoftRst}(\mathcal{H}_j) = \mathrm{SoftRst}(\mathcal{H}_{j'}). \tag{24}$$

   Further, since j, j' are hard interventions,

   $$\mathrm{Rst}(\mathcal{H}_j) = \mathrm{Rst}(\mathcal{H}_{j'}). \tag{25}$$

   In other terms, the high-level interventions must be distinct but their restrictions must be equal. When j, j' are hard interventions this is not possible, thus ω is unique. ■

2. $\omega(\varepsilon) = \varepsilon$,

   Proof Given an empty intervention , its soft restriction on the lower-level model contains all possible endogenous values. Formally,

   $$\mathrm{SoftRst}(\mathcal{M}_\varepsilon) = \mathrm{Val}(\boldsymbol{X}). \tag{26}$$

   Since τ is surjective, it also holds that

   $$\begin{aligned}\mathrm{SoftRst}(\mathcal{H}_j) &= \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_\varepsilon)\} \\ &= \mathrm{Val}(\boldsymbol{Y}).\end{aligned} \tag{27}$$

   In the same way, the empty restriction $\mathrm{Rst}(\mathcal{H}_\varepsilon)$ on the high-level model coincides with its domain. So, $\mathrm{SoftRst}(\mathcal{H}_\varepsilon) = \mathrm{Val}(\boldsymbol{Y})$ implies

   $$\mathrm{SoftRst}(\mathcal{H}_\varepsilon) = \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_\varepsilon)\} \tag{28}$$

   Therefore, by definition of low soft causal abstraction, it holds that $\omega(\varepsilon) = \varepsilon$. ■

3. $i_1 \preceq i_2 \implies \omega(i_1) \preceq \omega(i_2)$,

   Proof By ordering interventions according to their soft restriction, it follows that

$$\begin{aligned} i_1 \preceq i_2 &\implies \mathrm{SoftRst}(\mathcal{L}_{i_1}) \supseteq \mathrm{SoftRst}(\mathcal{L}_{i_2}) \\ &\implies \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_{i_1})\} \supseteq \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_{i_1})\} \\ &\implies \mathrm{SoftRst}(\mathcal{H}_{\omega(i_1)}) \supseteq \mathrm{SoftRst}(\mathcal{L}_{\omega(i_2)}) \\ &\implies \omega(i_1) \preceq \omega(i_2). \end{aligned} \tag{29}$$

■

4. There exists a subset of admissible hard interventions $\tilde{\mathcal{I}}$ such that $\mathcal{H}$ τ-abstracts $\mathcal{L}$ given a subset of high-level interventions $\tilde{\mathcal{J}}$.

   Proof Let $\tilde{\mathcal{I}}$ be the subset of hard interventions in $\mathcal{I}$ and $\tilde{\mathcal{J}} = \{\omega(i) \mid i \in \tilde{\mathcal{I}}\}$. Then, since $\mathcal{H}$ low soft τ-abstracts $\mathcal{L}$ for each intervention $i \in \mathcal{I}$ with there exists a partial function $\omega' \colon \tilde{\mathcal{I}} \to \tilde{\mathcal{J}}$ such that for any intervention $i \in \tilde{I}$ it holds

$$\mathrm{SoftRst}(\mathcal{H}_{\omega(i)}) = \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_i)\} \tag{30}$$

   and for every lower-level exogenous setting $\boldsymbol{e} \in \boldsymbol{E}$, it holds

$$\tau_{\boldsymbol{Y}}(\mathcal{L}_i(\boldsymbol{e})) = \mathcal{H}_{\omega(i)}(\tau_{\boldsymbol{U}}(\boldsymbol{e})). \tag{31}$$

   Thus, $\mathcal{H}$ τ-abstracts $\mathcal{L}$ given the interventions sets $\tilde{\mathcal{I}}, \tilde{\mathcal{J}}$. ■

A.3. Ambiguity in Lower Soft Abstraction

Theorem 10: Let $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ be an SCM with admissible interventions $\mathcal{J}$. Given a variable $V \in \boldsymbol{Y}$, if $\mathcal{J}$ contains two distinct interventions $j = (V \leftarrow g)$ and $j' = (V \leftarrow g')$ s.t.

1. $\forall \boldsymbol{u} \in \mathrm{Val}(\boldsymbol{U})\colon\ \mathrm{Proj}(\mathcal{H}_j(\boldsymbol{u}), V) = \mathrm{Proj}(\mathcal{H}_{j'}(\boldsymbol{u}), V)$,
2. $\mathrm{Image}(g) = \mathrm{Image}(g')$,

then, for any causal model $\mathcal{L}$ with admissible interventions $\mathcal{I}$, whenever $\mathcal{H}$ low soft τ-abstracts $\mathcal{L}$, the corresponding function $\omega \colon \mathcal{I} \to \mathcal{J}$ is not uniquely defined.

Proof Since $\mathcal{H}$ lower soft τ-abstracts $\mathcal{L}$, the function ω is surjective. Therefore, since $j_1, j_2 \in \mathcal{J}$, there must exist two lower-level intervention i_1, i_2 such that $\omega(i_1) = j_1$ and $\omega(i_2) = j_2$. Therefore, it holds that

$$\mathrm{SoftRst}(\mathcal{M}_{j_1}) = \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_{i_1})\} \tag{32}$$

$$\mathrm{SoftRst}(M_{j_2}) = \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_{i_2})\} \tag{33}$$

$$\tau_{\boldsymbol{Y}}(\mathcal{L}_{i_1}(\boldsymbol{e})) = \mathcal{H}_{j_1}(\tau_{\boldsymbol{U}}(\boldsymbol{e})) \tag{34}$$

$$\tau_{\boldsymbol{Y}}(\mathcal{L}_{i_2}(\boldsymbol{e})) = \mathcal{H}_{j_2}(\tau_{\boldsymbol{U}}(\boldsymbol{e})). \tag{35}$$

Since the images $\mathrm{Image}(g) = \mathrm{Image}(g')$ of the two high-level function coincide, it also holds that

$$\mathrm{SoftRst}(\mathcal{M}_{j_1}) = \mathrm{SoftRst}(M_{j_2}), \tag{36}$$

Furthermore, since g, g' output the same values given every possible configuration of their parents, for every low-level exogenous setting $\boldsymbol{e} \in \boldsymbol{E}$, it holds that

$$\boldsymbol{G}^{\mathcal{H}_{j_1}}(\tau_{\boldsymbol{U}}(\boldsymbol{e})) = \boldsymbol{G}^{\mathcal{H}_{j_2}}(\tau_{\boldsymbol{U}}(\boldsymbol{e})). \tag{37}$$

Consequently, while distinct, the two functions are equivalent given the constraints imposed by the definition of low soft abstraction. Therefore, it exist a function ω' such that

$$\omega'(i) = \begin{cases} \omega(i_1) & i = i_2 \\ \omega(i_2) & i = i_1 \\ \omega(i) & \text{otherwise} \end{cases}, \tag{38}$$

and $\mathcal{H}$ still lower soft τ-abstracts $\mathcal{L}$. Consequently, ω is not unique. ■

A.4. Soft Abstraction

Definition 11: Let $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$ and $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ be two SCMs, $\mathcal{I}$ and $\mathcal{J}$ be their respective sets of admissible interventions, $\tau_{\boldsymbol{Y}} \colon \mathrm{Val}(\boldsymbol{X}) \to \mathrm{Val}(\boldsymbol{Y})$ be a surjective function between endogenous settings, and $\tau_{\boldsymbol{U}} \colon \mathrm{Val}(\boldsymbol{E}) \to \mathrm{Val}(\boldsymbol{U})$ be a surjective function between exogenous settings. The model $\mathcal{H}$ is a soft τ-abstraction of $\mathcal{L}$ whenever it exists a surjective function $\omega \colon \mathcal{I} \to \mathcal{J}$ such that, for any intervention $i \in \mathcal{I}$ it holds

$$\mathrm{SoftRst}(\mathcal{H}_{\omega(i)}) = \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_i)\} \tag{39}$$

and for every lower-level exogenous $\boldsymbol{e} \in \boldsymbol{E}$ and endogenous $\boldsymbol{x} \in \boldsymbol{X}$ setting, it holds

$$\tau_{\boldsymbol{Y}}(\boldsymbol{F}^i(\boldsymbol{x}, \boldsymbol{e})) = \boldsymbol{G}^{\omega(i)}(\tau_{\boldsymbol{Y}}(\boldsymbol{x}), \tau_{\boldsymbol{U}}(\boldsymbol{e})). \tag{40}$$

Given Definition 11 of Soft Abstraction, we prove that whenever $\mathcal{H}$ soft τ-abstracts $\mathcal{L}$, it holds that

1. ω is unique.

   Proof We wish to show that for any distinct surjective function $\omega' \neq \omega$, $\mathcal{H}$ is not a soft τ-abstraction of $\mathcal{L}$. Since ω and ω' are defined on $\mathcal{I}$, they are distinct if there exists at least one intervention $i \in \mathcal{I}$ such that $\omega(i) \neq \omega'(i)$. Thus, there must be two distinct high-level interventions $j = (\boldsymbol{V} \leftarrow \boldsymbol{g})$ and $j' = (\boldsymbol{W} \leftarrow \boldsymbol{g'})$ in $\mathcal{J}$ such that $\omega(i) = j$ and $\omega'(i) = j'$. Consequently, for every low-level exogenous setting $\boldsymbol{e} \in \boldsymbol{E}$ and endogenous setting $\boldsymbol{x} \in \boldsymbol{X}$, it must hold

$$\tau_{\boldsymbol{Y}}(\boldsymbol{F}^i(\boldsymbol{x}, \boldsymbol{e})) = \boldsymbol{G}^{j}(\tau_{\boldsymbol{Y}}(\boldsymbol{x}), \tau_{\boldsymbol{U}}(\boldsymbol{e})) \tag{41}$$

$$\tau_{\boldsymbol{Y}}(\boldsymbol{F}^i(\boldsymbol{x}, \boldsymbol{e})) = \boldsymbol{G}^{j'}(\tau_{\boldsymbol{Y}}(\boldsymbol{x}), \tau_{\boldsymbol{U}}(\boldsymbol{e})). \tag{42}$$

   Therefore, it must also hold

$$G_Y^{j}(\tau_{\boldsymbol{Y}}(\boldsymbol{x}), \tau_{\boldsymbol{U}}(\boldsymbol{e})) = G_Y^{j'}(\tau_{\boldsymbol{Y}}(\boldsymbol{x}), \tau_{\boldsymbol{U}}(\boldsymbol{e})). \tag{43}$$

   for any high-level endogenous variable $Y \in \boldsymbol{Y}$, including those in $\boldsymbol{V}$, $\boldsymbol{W}$, or both. Consequently, since the output of the intervened functions match for any possible input, the high-level interventions j, j' are not distinct and ω is unique. ■

2. $\omega(\varepsilon) = \varepsilon$.

   Proof Same proof to Point 2 in Appendix A.2. ■

3. $i_1 \preceq i_2 \implies \omega(i_1) \preceq \omega(i_2)$.

   Proof Same proof to Point 3 in Appendix A.2. ■

4. Given the subset of admissible high-level hard interventions $\tilde{\mathcal{J}}$, $\mathcal{H}$ low soft τ-abstracts $\mathcal{L}$.

   Proof Since $\mathcal{H}$ soft τ-abstracts $\mathcal{L}$, and $\tilde{\mathcal{J}} \subseteq \mathcal{J}$, for any $j \in \tilde{\mathcal{J}}$ there exists $i \in \mathcal{I}$ such that $\omega(i) = j$. Therefore, we define

$$\tilde{\mathcal{I}} = \{i \in \mathcal{I} \mid \omega(i) \in \tilde{\mathcal{J}}\}. \tag{44}$$

   Consequently, for any $i \in \tilde{\mathcal{I}}$ it holds that

$$\mathrm{SoftRst}(\mathcal{H}_{\omega(i)}) = \{\tau_{\boldsymbol{Y}}(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathrm{SoftRst}(\mathcal{L}_i)\} \tag{45}$$

   and for every lower-level exogenous $\boldsymbol{e} \in \boldsymbol{E}$ and endogenous $\boldsymbol{x} \in \boldsymbol{X}$ setting, it holds

$$\tau_{\boldsymbol{Y}}(\boldsymbol{F}^i(\boldsymbol{x}, \boldsymbol{e})) = \boldsymbol{G}^{\omega(i)}(\tau_{\boldsymbol{Y}}(\boldsymbol{x}), \tau_{\boldsymbol{U}}(\boldsymbol{e})). \tag{46}$$

   Therefore, it also holds

$$\tau_{\boldsymbol{Y}}(\mathcal{L}^i(\boldsymbol{e})) = \mathcal{H}^{\omega(i)}(\tau_{\boldsymbol{U}}(\boldsymbol{e})). \tag{47}$$

   for any $\boldsymbol{e} \in \boldsymbol{E}$. Thus, $\mathcal{H}$ low soft τ-abstracts $\mathcal{L}$. ■

5. Given the subset of admissible low and high-level hard interventions $\tilde{\mathcal{I}}, \tilde{\mathcal{J}}$, $\mathcal{H}$ τ-abstracts $\mathcal{L}$.

   Proof The proof is equal to the previous, except we define the set of low-level admissible interventions as $\tilde{\mathcal{I}} = \{i \in \mathcal{I} \mid \omega(i) \in \mathcal{J} \wedge i \text{ is "hard"} \wedge \omega(i) \text{ is "hard}\}$, and consequently $\tilde{\mathcal{J}} = \{\omega(i) \mid i \in \tilde{\mathcal{I}}\}$. ■

A.5. Explicit Intervention Transformation

Theorem 14: Let $\mathcal{L} = (\boldsymbol{X}, \boldsymbol{E}, \boldsymbol{F}, \mathbb{P}_{\boldsymbol{E}})$ and $\mathcal{H} = (\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{G}, \mathbb{P}_{\boldsymbol{U}})$ be two SCMs and $\mathcal{I}$ and $\mathcal{J}$ their corresponding sets of soft interventions. Whenever $\mathcal{H}$ soft τ-abstracts $\mathcal{L}$ with alignment Π, ω maps the a low-level intervention $i = (\boldsymbol{V} \leftarrow \mathbf{f})$ to an intervention $j = (\boldsymbol{W} \leftarrow \mathbf{g})$ on the high-level variables $\boldsymbol{W} \subseteq \{Y \mid \boldsymbol{V} \cap \Pi(Y) \neq \emptyset\}$ such that

$$g_Y(\boldsymbol{y}, \boldsymbol{u}) = \tau_Y(F^i_{\Pi(Y)}(\tau^{-1}_{\boldsymbol{Y}}(\boldsymbol{y}), \tau^{-1}_{\boldsymbol{U}}(\boldsymbol{u}))), \tag{48}$$

for any variable $Y \in \boldsymbol{W}$ and any partial inverse τ^{-1}, where g_Y depends on its parents $\mathrm{Pa}(Y)$ only.
Proof Since $\mathcal{H}$ soft τ-abstracts $\mathcal{L}$, for each exogenous setting $\boldsymbol{e} \in \mathrm{Val}(\boldsymbol{E})$, for each endogenous setting $\boldsymbol{x} \in \mathrm{Val}(\boldsymbol{X})$, and each intervention $i \in \mathcal{I}$, it holds

$$\tau_{\boldsymbol{Y}}(F^i(\boldsymbol{x}, \boldsymbol{e})) = G^{\omega(i)}(\tau_{\boldsymbol{X}}(\boldsymbol{x}), \tau_{\boldsymbol{E}}(\boldsymbol{e})). \tag{49}$$

To simplify the notation, we denote as $\boldsymbol{x} \in \mathrm{Val}(\boldsymbol{X} \cup \boldsymbol{E})$ any combination of exogenous and endogenous settings. Therefore, without loss of generality, the previous equation is equivalent to

$$\tau_{\boldsymbol{X}}(F^i(\boldsymbol{x})) = G^{\omega(i)}(\tau(\boldsymbol{x})). \tag{50}$$

Given the mapping Π, the equation is also equivalent to

$$\tau_Y(F^i_{\Pi(Y)}(\boldsymbol{x})) = G^{\omega(i)}_Y(\tau(\boldsymbol{x})). \tag{51}$$

for any $Y \in \boldsymbol{Y}$. In the following, we prove that this equation is satisfied by our explicit definition of ω for any intervention i and any exogenous and endogenous configuration $\boldsymbol{x} \in \boldsymbol{X} \cup \boldsymbol{E}$.

Given any $\boldsymbol{x}$, applying the partial inverse $\tau^{-1}(\boldsymbol{x})$ returns a possibly different value $\boldsymbol{x'}$ such that $\tau(\boldsymbol{x}) = \tau(\boldsymbol{x'})$. Whenever $\mathcal{H}$ soft τ-abstracts $\mathcal{L}$, for any $\boldsymbol{x'}$ possibly different from $\boldsymbol{x}$, it holds that

$$\begin{aligned} \tau_{\mathrm{Pa}(Y)}(\boldsymbol{x}) &= \tau_{\mathrm{Pa}(Y)}(\boldsymbol{x'}) \\ \implies G^{\omega(i)}_Y(\tau(\boldsymbol{x})) &= G^{\omega(i)}_Y(\tau(\boldsymbol{x'})), \end{aligned} \tag{52}$$

given that the structural equation $G^{\omega(i)}_Y$ depends only on parents $\mathrm{Pa}(Y)$. Consequently,

$$\begin{aligned} G^{\omega(i)}_Y(\tau(\boldsymbol{x})) &= G^{\omega(i)}_Y(\tau(\boldsymbol{x'})) \\ \implies \tau_Y(F^i_{\Pi(Y)}(\tau(\boldsymbol{x}))) &= \tau_Y(F^i_{\Pi(Y)}(\tau(\boldsymbol{x'}))). \end{aligned} \tag{53}$$

Therefore, whenever $\mathcal{H}$ soft τ-abstracts $\mathcal{L}$, applying the explicit form from Equation 15, results in

$$G^{\omega(i)}_Y(\tau(\boldsymbol{x})) = \tau_Y(F^i_{\Pi(Y)}(\tau^{-1}(\tau(\boldsymbol{x})))) \tag{54}$$

$$= \tau_Y(F^i_{\Pi(Y)}(\boldsymbol{x'})))) \tag{55}$$

$$= \tau_Y(F^i_{\Pi(Y)}(\boldsymbol{x})))), \tag{56}$$

for any exogenous and endogenous configuration $\boldsymbol{x} \in \mathrm{Val}(\boldsymbol{X} \cup \boldsymbol{E})$. The application of the partial inverse depends independently on each variable Y, given the alignment Π. Furthermore, the step in Equation 56 depends only on the parents $\mathrm{Pa}(Y)$ of Y. Therefore, overall the intervened structural equation $G^{\omega(i)}_Y$ of Y, depends only on the parents $\mathrm{Pa}(Y)$.

Finally, we want to prove that intervening on a subset $\boldsymbol{V}$ of lower-level variables results *at most* in an high-level intervention on the variables $\boldsymbol{W} \subseteq \{Y \mid \boldsymbol{V} \cap \Pi(Y) \neq \emptyset\}$ whose clusters intersect with $\boldsymbol{V}$. In fact, for any high-level variable Y such that $\Pi(Y) \cap \boldsymbol{V} = \emptyset$, it holds that

$$G^{\omega(i)}_Y(\tau(\boldsymbol{x})) = \tau_Y(F^i_{\Pi(Y)}(\boldsymbol{x})))) \tag{57}$$

$$= \tau_Y(F_{\Pi(Y)}(\boldsymbol{x})))) \tag{58}$$

$$= G_Y(\tau(\boldsymbol{x})). \tag{59}$$

Therefore, applying ω results in the original structural equation G_Y given an intervention i not on in the cluster $\Pi(Y)$. Consequently, the variable Y is not intervened by $\omega(i)$ and $Y \notin \boldsymbol{W}$.

■