Abstracting Causal Models

Sander Beckers
Dept. of Philosophy and Religious Studies
Utrecht University
Utrecht, Netherlands
srekcebrednas@gmail.com

Joseph Y. Halpern
Dept. of Computer Science
Cornell University
Ithaca, NY 14853
halpern@cs.cornell.edu

Abstract

We consider a sequence of successively more restrictive definitions of abstraction for causal models, starting with a notion introduced by Rubenstein et al. (2017) called *exact transformation* that applies to probabilistic causal models, moving to a notion of *uniform transformation* that applies to deterministic causal models and does not allow differences to be hidden by the "right" choice of distribution, and then to *abstraction*, where the interventions of interest are determined by the map from low-level states to high-level states, and *strong abstraction*, which takes more seriously all potential interventions in a model, not just the allowed interventions. We show that procedures for combining micro-variables into macro-variables are instances of our notion of strong abstraction, as are all the examples considered by Rubenstein et al.

1 Introduction

We can and typically do analyze problems at different levels of abstraction. For example, we can try to understand human behavior by thinking at the level of neurons firing in the brain or at the level of beliefs, desires, and intentions. A political scientist might try to understand an election in terms of individual voters or in terms of the behavior of groups such as midwestern blue-collar workers. Since, in these analyses, we are typically interested in the *causal* connections between variables, it seems reasonable to model the various levels of abstraction using causal models (Halpern 2016; Pearl 2000). The question then arises whether a high-level "macro" causal model (e.g., one that considers beliefs, desires, and intentions) is a faithful abstraction of a low-level "micro" model (e.g., one that describes things at the neuronal level). What should this even mean?

Perhaps the most common way to approach the question of abstraction is to cluster "micro-variables" in the low-level model into a single "macro-variable" in the high-level model (Chalupka, Eberhardt, and Perona 2015; 2016; Iwasaki and Simon 1994). Of course, one has to be careful to do this in a way that preserves the causal relationships in the low-level model. For example, we do not want to cluster variables X, Y, and Z into a single variable $X + Y + Z$ if different settings (x, y, z) and (x', y', z') such that $x + y + z = x' + y' + z'$ lead to different outcomes.

Rubenstein et al. (2017) (RW$^+$ from now on) provided an arguably more general approach to abstraction. They defined a notion of an *exact transformation* between two causal models. They suggest that if there is an exact transformation τ from causal model M_1 to M_2, then we should think of M_2 as an abstraction of M_1, so that M_2 is the high-level model and M_1 is the low-level model.

Abstraction almost by definition involves ignoring inessential differences. So it seems that RW+ would want to claim that if there exists an exact transformation from M_1 and M_2, then M_2 and M_1 are the same, except for "inessential differences". This leads to the obvious question: what counts as an inessential difference? Of course, this is to some extent in the eye of the beholder, and may well depend on the application. Nevertheless we claim that the notion of "inessential difference" implicitly encoded in the definition of exact transformation is far too broad. As we show by example, there are models that we would view as significantly different that are related by exact transformations. There are two reasons for this. The first is that, because RW$^+$ consider probabilistic causal models, some differences that are intuitively significant are overlooked by considering just the right distributions. Second, besides a function that maps low-level states to high-level states, RW$^+$ define a separate mapping of interventions that can mask what we view as essential differences between the interventions allowed at the low level and the high level.

In this paper, we consider a sequence of successively more restrictive definitions of abstraction, starting with the RW$^+$ notion of exact transformation, moving to a notion of *uniform transformation* that applies to deterministic causal models and does not allow differences to be hidden by the "right" choice of distribution, and then to *abstraction*, where the mapping between the interventions is determined by the mapping from low-level states to high-level states, and *strong abstraction*, which takes more seriously all potential interventions in a model, not just the allowed interventions. Finally, we define *constructive abstraction*, which is the special case of strong abstraction where the mapping from low-level states to high-level states partitions the low-level variables and maps each cell to a unique high-level variable. As we show, procedures for combining micro-variables into macro-variables are instances of constructive abstraction, as are all the other examples considered by RW$^+$. While we

view constructive abstraction as the notion that is likely to be the most useful in practice, as we show by example, the weaker notions of strong abstraction and abstraction are of interest as well.

Not surprisingly, the idea of abstracting complicated low-level models to simpler high-level models that in some sense act the same way has also been considered in other settings; see, for example, (Binahashemi, de Giacomo, and Lespérance 2017). While we are trying to capture these intuitions as well, considering a setting that involves causality adds new subtleties.

2 Probabilistic causal models: a review

In this section we review the definition of causal models. Much of the discussion is taken from (Halpern 2016).

Definition 2.1: A signature $\mathcal{S}$ is a tuple $(\mathcal{U}, \mathcal{V}, \mathcal{R})$, where $\mathcal{U}$ is a set of *exogenous* variables (intuitively, variables that represent factors outside the control of the model), $\mathcal{V}$ is a set of *endogenous* variables (intuitively, variables whose values are ultimately determined by the values of the endogenous variables), and $\mathcal{R}$, a function that associates with every variable $Y \in \mathcal{U} \cup \mathcal{V}$ a nonempty set $\mathcal{R}(Y)$ of possible values for Y (i.e., the set of values over which Y *ranges*). If $\vec{X} = (X_1, \ldots, X_n)$, $\mathcal{R}(\vec{X})$ denotes the crossproduct $\mathcal{R}(X_1) \times \cdots \times \mathcal{R}(X_n)$. ∎

Definition 2.2: A *basic causal model* M is a pair $(\mathcal{S}, \mathcal{F})$, where $\mathcal{S}$ is a signature and $\mathcal{F}$ defines a function that associates with each endogenous variable X a *structural equation* F_X giving the value of X in terms of the values of other endogenous and exogenous variables (discussed in more detail below). A *causal model* M is a tuple $(\mathcal{S}, \mathcal{F}, \mathcal{I})$, where $(\mathcal{S}, \mathcal{F})$ is a basic causal model and $\mathcal{I}$ is a set of *allowed interventions* (also discussed in more detail below). ∎

Formally, the equation F_X maps $\mathcal{R}(\mathcal{U} \cup \mathcal{V} - \{X\})$ to $\mathcal{R}(X)$, so F_X determines the value of X, given the values of all the other variables in $\mathcal{U} \cup \mathcal{V}$. Note that there are no functions associated with exogenous variables; their values are determined outside the model. We call a setting $\vec{u}$ of values of exogenous variables a *context*.

The value of X may depend on the values of only a few other variables. Y *depends on X in context $\vec{u}$* if there is some setting of the endogenous variables other than X and Y such that if the exogenous variables have value $\vec{u}$, then varying the value of X in that context results in a variation in the value of Y; that is, there is a setting $\vec{z}$ of the endogenous variables other than X and Y and values x and x' of X such that $F_Y(x, \vec{z}, \vec{u}) \neq F_Y(x', \vec{z}, \vec{u})$.

In this paper we restrict attention to *recursive* (or *acyclic*) models, that is, models where, for each context $\vec{u}$, there is a partial order $\prec_{\vec{u}}$ on variables such that if $X \prec_{\vec{u}} Y$, then Y depends on X in context $\vec{u}$. In a recursive model, given a context $\vec{u}$, the values of all the remaining variables are determined (we can just solve for the value of the variables in the order given by $\prec_{\vec{u}}$). A model is *strongly* recursive if the partial order $\prec_{\vec{u}}$ is independent of $\vec{u}$; that is, there is a partial order $\prec$ such that $\prec = \prec_{\vec{u}}$ for all contexts $\vec{u}$. In a strongly

recursive model, we often write the equation for an endogenous variable as $X = f(\vec{Y})$; this denotes that the value of X depends only on the values of the variables in $\vec{Y}$, and the connection is given by f. For example, we might have $X = Y + U$. [1]

An *intervention* has the form $\vec{X} \leftarrow \vec{x}$, where $\vec{X}$ is a set of endogenous variables. Intuitively, this means that the values of the variables in $\vec{X}$ are set to $\vec{x}$. The structural equations define what happens in the presence of external interventions. Setting the value of some variables $\vec{X}$ to $\vec{x}$ in a causal model $M = (\mathcal{S}, \mathcal{F})$ results in a new causal model, denoted $M_{\vec{X} \leftarrow \vec{x}}$, which is identical to M, except that $\mathcal{F}$ is replaced by $\mathcal{F}^{\vec{X} \leftarrow \vec{x}}$: for each variable $Y \notin \vec{X}$, $F_Y^{\vec{X} \leftarrow \vec{x}} = F_Y$ (i.e., the equation for Y is unchanged), while for each X' in $\vec{X}$, the equation $F_{X'}$ for is replaced by $X' = x'$ (where x' is the value in $\vec{x}$ corresponding to $\vec{x}$).

The set $\mathcal{I}$ of interventions can be viewed as the set of interventions that we care about for some reason or other. For example, it might consist of the interventions that involve variables and values that are under our control. In (Halpern and Pearl 2005; Halpern 2016), only basic causal models are considered (and are called causal models). RW$^+$ added the set of allowed interventions to the model. We consider allowed interventions as well, since it seems useful when considering abstractions to describe the set of interventions of interest. We sometimes write a causal model $M = (\mathcal{S}, \mathcal{F}, \mathcal{I})$ as $(M', \mathcal{I})$, where M' is the basic causal model $(\mathcal{S}, \mathcal{F})$, if we want to emphasize the role of the set of interventions.

Given a signature $\mathcal{S} = (\mathcal{U}, \mathcal{V}, \mathcal{R})$, a *primitive event* is a formula of the form $X = x$, for $X \in \mathcal{V}$ and $x \in \mathcal{R}(X)$. A *causal formula (over $\mathcal{S}$)* is one of the form $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$, where

- φ is a Boolean combination of primitive events,

- $Y_1, \ldots, Y_k$ are distinct variables in $\mathcal{V}$, and

- $y_i \in \mathcal{R}(Y_i)$.

Such a formula is abbreviated as $[\vec{Y} \leftarrow \vec{y}]\varphi$. The special case where $k = 0$ is abbreviated as φ. Intuitively, $[Y_1 \leftarrow y_1, \ldots, Y_k \leftarrow y_k]\varphi$ says that φ would hold if Y_i were set to y_i, for $i = 1, \ldots, k$.

A causal formula ψ is true or false in a causal model, given a context. As usual, we write $(M, \vec{u}) \models \psi$ if the causal formula ψ is true in causal model M given context $\vec{u}$. The $\models$ relation is defined inductively. $(M, \vec{u}) \models X = x$ if the variable X has value x in the unique (since we are dealing with recursive models) solution to the equations in M in context $\vec{u}$ (i.e., the unique vector of values that simultaneously satisfies all equations in M with the variables in $\mathcal{U}$ set to $\vec{u}$). The

---

[1]RW$^+$ do not restrict to acyclic models. Rather, they make the weaker restriction that, for every setting of the causal variables, with probability 1, there is a unique solution to the equations. In the deterministic setting, the analogous restriction would be to consider causal models where there is a unique solution to all equations. None of our definitions or results changes if we allow this more general class of models. We have restricted to recursive models only to simplify the exposition.

truth of conjunctions and negations is defined in the standard way. Finally, $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}]\varphi$ if $(M_{\vec{Y} \leftarrow y}, \vec{u}) \models \varphi$.

To simplify notation, we sometimes write $M(\vec{u})$ to denote the unique element of $\mathcal{R}(\mathcal{V})$ such that $(M, \vec{u}) \models \mathcal{V} = \vec{v}$. Similarly, given an intervention $\vec{Y} \leftarrow \vec{y}$, $M(\vec{u}, \vec{Y} \leftarrow \vec{y})$ denotes the unique element of $\mathcal{R}(\mathcal{V})$ such that $(M, \vec{u}) \models [\vec{Y} \leftarrow \vec{y}](\mathcal{V} = \vec{v})$.

A *probabilistic causal model* $M = (\mathcal{S}, \mathcal{F}, \mathcal{I}, \mathrm{Pr})$ is just a causal model together with a probability Pr on contexts. We often abuse notation slightly and denote the probabilistic causal model $(\mathcal{S}, \mathcal{F}, \mathcal{I}, \mathrm{Pr})$ as (M, Pr), where M is the underlying deterministic causal model $(\mathcal{S}, \mathcal{F}, \mathcal{I})$.

RW$^+$ worked with probabilistic causal models, but added one more feature and made a restrictive assumption. They consider models M that place a partial order $\prec_M$ on interventions. However, they (and we) consider only what they call the *natural* partial order, where $(\vec{X} \leftarrow \vec{x}) \prec_M (\vec{X}' \leftarrow \vec{x}')$ if $\vec{X}$ is a subset of $\vec{X}'$ and $\vec{x}$ is the corresponding subset of $\vec{x}'$, so we do not explicitly introduce the partial order as a component of the model here. In addition, RW$^+$ assume that for each endogenous variable X, there is a unique exogenous variable U_X such that U_X is the only exogenous variable on whose value X depends, and $U_X \neq U_Y$ if $X \neq Y$. We say that a causal model has *unique exogenous variables (uev)* if this is the case.

Assuming that a causal model has uev makes sense if we think of U_X as the noise variable corresponding to X. However, this assumption is not always appropriate (e.g., if we take the temperature to be exogenous, and temperature can affect a number of endogenous variables). Not surprisingly, in (non-probabilistic) causal models, assuming uev entails a significant loss of generality. In particular, we cannot express the correlation in values between two endogenous variables due to being affected by a common exogenous variable. However, the uev assumption can be made essentially without loss of generality in probabilistic causal models, as the lemma below shows.

Definition 2.3 : Two probabilistic causal models $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F}, \mathcal{I}, \mathrm{Pr})$ and $M' = ((\mathcal{U}', \mathcal{V}', \mathcal{R}'), \mathcal{F}', \mathcal{I}', \mathrm{Pr}')$ are *equivalent*, written $M \sim M'$, if $\mathcal{V} = \mathcal{V}'$, $\mathcal{R}(Y) = \mathcal{R}'(Y)$ for all $Y \in \mathcal{V}$, $\mathcal{I} = \mathcal{I}'$, and all causal formulas have the same probability of being true in both M and M'; that is, for all causal formulas φ, we have $\mathrm{Pr}(\{\vec{u} \in \mathcal{R}(\mathcal{U}) : (M, \vec{u}) \models \varphi\}) = \mathrm{Pr}'(\{\vec{u}' \in \mathcal{R}(\mathcal{U}') : (M', \vec{u}') \models \varphi\})$. ∎

Lemma 2.4: *Given a probabilistic causal model M, there is a probabilistic causal model M' with uev such that $M \sim M'$.*[2]

All the models that we consider in our examples have uev. Whatever problems there are with the RW$^+$ notions, they do not arise from the assumption that models have uev.

3 From exact transformations to abstractions

In this section, we review the RW$^+$ definition, point out some problems with it, and then consider a sequence of strengthenings of the definition.

---

[2]Proofs can be found in the appendix.

3.1 Exact transformations

We need some preliminary definitions. First observe that, given a probabilistic model $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F}, \mathcal{I}, \mathrm{Pr})$, the probability Pr on $\mathcal{R}(\mathcal{U})$ can also be viewed as a probability on $\mathcal{R}(\mathcal{V})$ (since each context in $\mathcal{R}(\mathcal{U})$ determines a unique setting of the variables in $\mathcal{V}$); more precisely,

$$\mathrm{Pr}(\vec{v}) = \mathrm{Pr}(\{\vec{u} : M(\vec{u}) = \vec{v}\}).$$

In the sequel, we freely view Pr as a distribution on both $\mathcal{R}(\mathcal{U})$ and $\mathcal{R}(\mathcal{V})$; the context should make clear which we intend. Each intervention $\vec{X} \leftarrow \vec{x}$ also induces a probability $\mathrm{Pr}^{\vec{X} \leftarrow \vec{x}}$ on $\mathcal{R}(\mathcal{V})$ in the obvious way:

$$\mathrm{Pr}^{\vec{X} \leftarrow \vec{x}}(\vec{v}) = \mathrm{Pr}(\{\vec{u} : M(\vec{u}, \vec{X} \leftarrow \vec{x}) = \vec{v}\}).$$

One last piece of notation: We are interested in when a high-level model is an abstraction of a low-level model. In the sequel, we always use $M_L = ((\mathcal{U}_L, \mathcal{V}_L, \mathcal{R}_L), \mathcal{F}_L, \mathcal{I}_L)$ and $M_H = ((\mathcal{U}_H, \mathcal{V}_H, \mathcal{R}_H), \mathcal{F}_H, \mathcal{I}_H)$ to denote deterministic causal models (where the L and H stand for *low level* and *high level*, respectively). We write (M, Pr) to denote a probabilistic causal model that extends M.

With this background, we can give the RW$^+$ definition of exact transformation. Although the definition was given for probabilistic causal models that satisfy uev, it makes sense for arbitrary probabilistic causal models.

Definition 3.1: If (M_L, Pr_L) and (M_H, Pr_L) are probabilistic causal models, $\omega : \mathcal{I}_L \to \mathcal{I}_H$ is an order-preserving, surjective mapping (where ω is *order-preserving* if, for all interventions in $i_1, i_2 \in \mathcal{I}_L$ such that $i_1 \prec_{M_L} i_2$ according to the natural order, we have $\omega(i_i) \prec_{M_H} \omega(i_2)$), and $\tau : \mathcal{R}_L(\mathcal{V}_L) \to \mathcal{R}_H(\mathcal{V}_H)$, then (M_H, Pr_H) is an *exact (τ-ω)-transformation* of (M_L, Pr_L) if, for every intervention $\vec{Y} \leftarrow \vec{y} \in \mathcal{I}_L$, we have

$$\mathrm{Pr}_H^{\omega(\vec{Y} \leftarrow \vec{y})} = \tau(\mathrm{Pr}_L^{\vec{Y} \leftarrow \vec{y}}), \tag{1}$$

where $\tau(\mathrm{Pr}_L)$ is the "pushforward" distribution on $\mathcal{R}(\mathcal{V}_H)$ determined by τ and Pr_L:

$$\tau(\mathrm{Pr}_L)(\vec{v}_H) = \mathrm{Pr}_L(\{\vec{v}_L : \tau(\vec{v}_L) = \vec{v}_H\}).$$

∎

The key point here is the requirement that $\mathrm{Pr}_H^{\omega(\vec{Y} \leftarrow \vec{y})} = \tau(\mathrm{Pr}_L^{\vec{Y} \leftarrow \vec{y}})$. Roughly speaking, it says that if you start from the low-level intervention $\vec{Y} \leftarrow \vec{y}$ and move up to the high-level model following two distinct routes, you end up at the same place.

The first route goes as follows. The intervention $\vec{Y} \leftarrow \vec{y}$ changes the probability distribution on low-level outcomes, giving rise to $\mathrm{Pr}_L^{\vec{Y} \leftarrow \vec{y}}$ (where an "outcome" is a setting of the endogenous variables). This distribution can be moved up to the high level by applying τ, giving $\tau(\mathrm{Pr}_L^{\vec{Y} \leftarrow \vec{y}})$, which is a distribution on high-level outcomes.

The second route goes as follows. From the low-level intervention $\vec{Y} \leftarrow \vec{y}$ we move up to a high-level intervention by applying ω, giving $\omega(\vec{Y} \leftarrow \vec{y})$. This intervention changes

the probability distribution on high-level outcomes, giving rise to $\mathrm{Pr}_H^{\omega(\vec{Y} \leftarrow \vec{y})}$. To be an exact transformation means that this distribution and the previous one are identical, for all interventions $\vec{Y} \leftarrow \vec{y}$.

Despite all the notation, we hope that the intuition is clear: the intervention $\vec{Y} \leftarrow \vec{y}$ acts the same way in the low-level model as the intervention $\omega(\vec{Y} \leftarrow \vec{y})$ does in the high level-model. (See RW$^+$ for more discussion and intuition.) The following example illustrates Definition 3.1.

Example 3.2: Consider a simple voting scenario where we have 99 voters who can either vote for or against a proposition. The campaign for the proposition can air some subset of two advertisements to try to influence how the voters vote. The low-level model is characterized by endogenous variables X_i, $i = 1, \ldots, 99$, A_1, A_2, and T, and exogenous variables U_i, $i = 1, \ldots, 101$. X_i denotes voter i's vote, so $X_i = 1$ if voter i votes for the proposition, and $X_i = 0$ if voter i votes against. A_i denotes whether add i is run, and T denotes the total number of votes for the proposition. U_i determines how voter i votes as a function of which ads are run for $i = 1, \ldots, 99$, while U_{100} and U_{101} determine A_1 and A_2, respectively.

We can cluster the voters into three groups: X_1–X_{33}, X_{34}–X_{66}, X_{67}–X_{99}. For example, the first group might represent older, wealthy voters; the second group might represent soccer moms; and the third group might represent young singles. Members of the same group are affected by the ads in the same way, meaning that $Pr(X_i = 1|A_1 = a_1 \wedge A_2 = a_2) = Pr(X_j = 1|A_1 = a_1 \wedge A_2 = a_2)$ for all a_1, a_2 and all i, j that belong to the same group. The high-level model replaces the variables $X_1, \ldots, X_{99}$ by variables G_1, G_2, and G_3, representing the sum of the votes of each group, it replaces $U_1, \ldots, U_{99}$ by U_1', U_2', U_3', and replaces T by a binary variable T' that just indicates who won. The only interventions allowed in the low-level model are interventions to the variables A_1 and A_2.

We now have an obvious map τ from $\mathcal{V}_L$ to $\mathcal{V}_H$ that maps a low-level state to a high-level state by taking G_1, G_2, and G_3 to be the total vote of the corresponding groups; the map ω is just the identity. Given a probability Pr_L on $\mathcal{U}_L$, there is an obvious probability Pr_H on $\mathcal{U}_H$ such that (M_H, Pr_H) is an exact transformation of (M_L, Pr_L). Note that it is critical here that we don't allow interventions on the individual variables X_i at the low level. For example, it is not clear to what high-level intervention ω should map the low-level intervention $X_3 \leftarrow 1$. ∎

RW$^+$ discuss three applications of exact transformations:

- a model from which some variables are marginalized;

- moving from the micro-level to the macro-level by aggregating groups of variables;

- and moving from a time-evolving dynamical process to a stationary equilibrium state.

We review the details of their second application here, just to show how it plays out in our framework.

Example 3.3: Let M_L be a causal model with endogenous variables $\vec{X} = \{X_i : 1 \le i \le n\}$ and $\vec{Y} = \{Y_i : 1 \le$

$i \le m\}$, and equations $X_i = U_i$ for $1 \le i \le n$ and $Y_i = \sum_n^{j=1} A_{ij} X_j + V_i$ for $1 \le i \le m$, where A is an $m \times n$ matrix and there exists an $a \in \mathbb{R}$ such that each column of the matrix A sums to a. Finally, the intervention set is $\mathcal{I}_L = \{\emptyset, \vec{X} \leftarrow \vec{x}, \vec{Y} \leftarrow \vec{y}, (\vec{X}, \vec{Y}) \leftarrow (\vec{x}, \vec{y}) : \vec{x} \in \mathbb{R}^n, \vec{y} \in \mathbb{R}^m\}$.

Let M_H be a model with endogenous variables $\bar{X}$ and $\bar{Y}$, equations $\bar{X} = \bar{U}$ and $\bar{Y} = \frac{a}{m}\bar{X} + \bar{V}$, and intervention set $\mathcal{I}_H = \{\emptyset, \bar{X} \leftarrow \bar{x}, \bar{Y} \leftarrow \bar{y}, (\bar{X}, \bar{Y}) \leftarrow (\bar{x}, \bar{y}) : \bar{x}, \bar{y} \in \mathbb{R}\}$.

Consider the following transformation that averages the X and Y variables:

$$\tau : \mathcal{R}(\mathcal{V}_L) \to \mathcal{R}(\mathcal{V}_H) = \mathbb{R}^2$$
$$(\vec{X}, \vec{Y}) \to (\tfrac{1}{n}\sum_n^{i=1} X_i, \tfrac{1}{m}\sum_m^{i=1} Y_i).$$

Given a probability $\mathrm{Pr}_{\mathcal{U}_L}$ on $\mathcal{U}_l$ (the contexts in the low-level model M_L), if we take $\mathrm{Pr}_{\bar{U}} = \mathrm{Pr}_{\mathcal{U}_L}(\frac{1}{n}\sum_n^{i=1} U_i)$ and $\mathrm{Pr}_{\bar{V}} = \mathrm{Pr}_{\mathcal{U}_L}(\frac{1}{m}\sum_m^{i=1} V_i)$, then M_H is an exact $(\tau\text{-}\omega)$-transformation of M_L for the obvious choice of ω. ∎

3.2 Uniform transformations

As the following example shows, much of the work to ensure that a transformation is an exact transformation can be done by choosing appropriate distributions Pr_L and Pr_H. This leads to cases where (M_H, Pr_H) is an exact transformation of (M_L, Pr_L) although it is hard to think of M_H as a high-level abstraction of M_L.

Example 3.4: For $i = \{1, 2\}$, let M_i be a deterministic causal model with signature $(\mathcal{U}_i, \mathcal{V}_i, \mathcal{R}_i)$; let $\vec{u}_i$ be a fixed context in M_i; let $\vec{v}_i \in \mathcal{R}_i(\mathcal{V}_i)$ be such that $(M_i, \vec{u}_i) \models \mathcal{V}_i = \vec{v}_i$; let $\mathcal{I}_i$ consist only of the empty intervention; let Pr_i put probability 1 on $\vec{u}_i$; let τ_i map all elements of $\mathcal{R}(\mathcal{V}_i)$ to $\vec{v}_{3-i}$; and let ω_i be the identity map from $\mathcal{I}_i$ to $\mathcal{I}_{3-i}$. Clearly (M_i, Pr_i) is an exact $(\tau_i\text{-}\omega_i)$-transformation of $(M_{3-i}, \mathrm{Pr}_{3-i})$. ∎

The fact that each of M_1 and M_2 is an exact transformation of the other, despite the fact that the models are completely unrelated, suggests to us that exact transformations are not capturing the essence of abstraction. Roughly speaking, what is happening here is that a high-level model M_H can be arbitrary in contexts that do not lead to settings $\vec{v}_H$ that have positive probability for some allowed low-level intervention. This means that if there are few allowed low-level interventions or few contexts with positive probability, then there are very few constraints on M_H. We end up with high-level models M_H that should not (in our view) count as abstractions of M_L. We can address this concern by strengthening the notion of exact transformation to require it to hold for *all* distributions Pr_L.

Definition 3.5: If M_L and M_H are deterministic causal models, ω is an order-preserving, surjective mapping $\omega : \mathcal{I}_L \to \mathcal{I}_H$, and $\tau : \mathcal{R}_L(\mathcal{V}_L) \to \mathcal{R}_H(\mathcal{V}_H)$, then M_H is a *uniform (τ-ω)-transformation* of M_L if, for all Pr_L, there exists Pr_H such that (M_H, Pr_H) is an exact $(\tau\text{-}\omega)$-transformation of (M_L, Pr_L). ∎

As we pointed out earlier, since RW$^+$ assume uev, the probability distribution in general might do a lot of work

to capture correlations between values of endogenous variables. It makes sense to consider arbitrary distributions if we drop the uev assumption (as in fact we do).

In Example 3.4 it is easy to see that neither M_1 nor M_2 is a uniform transformation of the other. On the other hand, in Example 3.2, we do have a uniform transformation.

Considering uniform transformations has other nice features. For one thing, it allows us to derive from τ a mapping $\tau_{\mathcal{U}}$ from $\mathcal{R}(\mathcal{U}_L)$ to $\mathcal{R}(\mathcal{U}_H)$ that "explains" how $\Pr_L$ and $\Pr_H$ are related. More precisely, not only do we know that, for the appropriate ω, for all distributions $\Pr_L$ there exists $\Pr_H$ such that $(M_H, \Pr_H)$ is an exact $(\tau\text{-}\omega)$-transformation of $(M_L, \Pr_L)$, we can take $\Pr_H$ to be $\tau_{\mathcal{U}}(\Pr_L)$ (i.e., the pushforward of $\Pr_L$ under $\tau_{\mathcal{U}}$).

Proposition 3.6: *If M_H is a uniform $(\tau\text{-}\omega)$-transformation of M_L, then there exists a function $\tau_{\mathcal{U}} : \mathcal{R}(\mathcal{U}_L) \to \mathcal{R}(\mathcal{U}_H)$ such that, for all distributions $\Pr_L$ on $\mathcal{R}(\mathcal{U}_L)$, $(M_H, \tau_{\mathcal{U}}(\Pr_L))$ is an exact $(\tau\text{-}\omega)$-transformation of $(M_L, \Pr_L)$.*

The next result provides a characterization of when M_H is a uniform $(\tau\text{-}\omega)$-transformation of M_L.

Definition 3.7: $\tau' : \mathcal{R}(\mathcal{U}_L) \to \mathcal{R}(\mathcal{U}_H)$ is *compatible* with $\tau : \mathcal{R}(\mathcal{V}_L) \to \mathcal{R}(\mathcal{V}_H)$ if, for all $\vec{Y} \leftarrow \vec{y} \in \mathcal{I}_L$ and $\vec{u}_L \in \mathcal{R}(\mathcal{U}_L)$,

$$\tau(M_L(\vec{u}_L, \vec{Y} \leftarrow \vec{y})) = M_H(\tau'(\vec{u}_L), \omega(\vec{Y} \leftarrow \vec{y})).$$

∎

Theorem 3.8: *Given causal models M_L and M_H, $\tau : \mathcal{R}(\mathcal{V}_L) \to \mathcal{R}(\mathcal{V}_H)$, and an order-preserving surjective function $\omega : \mathcal{I}(\mathcal{V}_L) \to \mathcal{I}(\mathcal{V}_H)$, the following are equivalent:*

(a) M_H is a uniform $(\tau\text{-}\omega)$-transformation of M_L;
(b) there exists a function $\tau_{\mathcal{U}} : \mathcal{R}(\mathcal{U}_L) \to \mathcal{R}(\mathcal{U}_H)$ compatible with τ.

It is easy to check that uniform transformations are closed under composition.

Theorem 3.9: *If M_H is a uniform $(\tau_1\text{-}\omega_1)$-transformation of M_I and M_I is a uniform $(\tau_2\text{-}\omega_2)$-transformation of M_L, then M_H is a uniform $((\tau_2 \circ \tau_1)\text{-}(\omega_2 \circ \omega_1))$-transformation of M_L.*

3.3 Abstraction

Although the notion of a uniform transformation deals with some of the problems we see with the RW^+ notion of exact transformation, it does not deal with all of them, as the following two examples show.

Example 3.10: Let M_1 and M_2 be deterministic causal models, both with endogenous binary variables X_1 and X_2 and corresponding binary exogenous variables U_1 and U_2. [3] In M_1, the equations are $X_1 = U_1$ and $X_2 = X_1$. (U_2 plays no role in the equations in M_1. We added it just to make M_2 a model that has uev and thus show that having uev is not an issue here.) In M_2, the equations are $X_1 = U_1$ and $X_2 = U_2$. The only allowed interventions in M_1 are

---

[3] A variable is binary if its range is $\{0, 1\}$.

$X_1 \leftarrow x_1$, for $x_1 \in \{0, 1\}$; the only allowed interventions in M_2 are $(X_1, X_2) \leftarrow (x_1, x_1)$, for $x_1 \in \{0, 1\}$. It is easy to see that M_1 is a uniform transformation of M_2 and that M_2 is a uniform transformation of M_1. If τ_{ij} and ω_{ij} are the maps showing that M_j is a uniform transformation of M_i, then we can take both τ_{12} and τ_{21} to be the identity, ω_{12} maps $X_1 \leftarrow x_1$ to $(X_1, X_2) \leftarrow (x_1, x_1)$, while ω_{21} maps $(X_1, X_2) \leftarrow (x_1, x_1)$ to $X_1 \leftarrow x_1$. But this does not match our intuition that if M_H is an abstraction of M_L, then M_H is a higher-level description of the situation than M_L. Whatever "higher-level description" means, we would expect that if M_L and M_H are different, then we should not have M_L and M_H being abstractions of each other. ∎

What is the problem here? If we just focus on these sets of allowed interventions, then there is in fact no problem. M_1 and M_2 do, in a sense, work the same way as far as these allowed interventions go. However, the mappings ω_{12} and ω_{21} seem to be in conflict with taking τ_{12} and τ_{21} to be the identity. Given that τ_{ij} is the identity mapping, we would expect ω_{ij} to also be the identity mapping. Why should ω_{12} map $X_1 \leftarrow 0$ to something other than $X_1 \leftarrow 0$ here? It is easy to see that if we take ω_{12} to also be the identity mapping then the problem disappears, as we no longer have uniform transformations between these two models. More generally, we define below a natural way in which a mapping τ on states induces a mapping ω on allowed interventions. But even when ω is well-behaved there exist counterintuitive examples of uniform transformations.

Example 3.11: Given a model M_3, let M_4 be a model that is like M_3 except that $\mathcal{U}_3$ has a new exogenous binary variable U^* and a new binary endogenous variable X^*. Modify the equations in M_3 so that U^* is the only parent of X^*, but X^* is the parent of every other endogenous variable in M_4 (and thus of every endogenous variable in M_3). Take $\mathcal{I}_4^* = \mathcal{I}_3$. If $X^* = 1$, then all equations in M_4 are identical to those in M_3. However, if $X^* = 0$, then all equations behave in some arbitrary way (the exact way they behave is irrelevant). Define $\tau : \mathcal{R}(\mathcal{V}_3) \to \mathcal{R}(\mathcal{V}_4)$ by taking $\tau(\vec{v}_L) = (\vec{v}_L, X^* = 1)$. We claim that M_4 is a uniform $(\tau\text{-}\omega)$-transformation of M_L, where ω is the identity. Given a distribution $\Pr_3$ on $\mathcal{R}(\mathcal{U}_3)$, define $\Pr_4$ so that its marginal on the variables in $\mathcal{U}_3$ is $\Pr_3$ and $\Pr_4(U^* = 0) = 0$. It is easy to see that $(M_4, \Pr_4)$ is an exact $(\tau\text{-}\omega)$-transformation of $(M_3, \Pr_3)$, regardless of how the equations in M_4 are defined if $X^* = 1$. ∎

What goes wrong in this example is that the high level is more detailed than the low level, contrary to what one expects of an abstraction. Concretely, introducing the extra variable X^* allows M_4 to capture a whole range of possibilities that have no counterpart whatsoever in M_3. That doesn't sound right (at least to us). We can fix this by simply demanding that our abstraction function τ be surjective.

Combining both observations, we define a natural way in which an abstraction function τ determines which sets of interventions should be allowed at the low level and the high level, and the mapping ω_τ between them.

Definition 3.12: Given a set $\mathcal{V}$ of endogenous variables,

$\vec{X} \subseteq \mathcal{V}$, and $\vec{x} \in \mathcal{R}(\vec{X})$, let

$$Rst(\mathcal{V}, \vec{x}) = \{\vec{v} \in \mathcal{R}(\mathcal{V}) : \vec{x} \text{ is the restriction of } \vec{v} \text{ to } \vec{X}\}.$$

Given $\tau : \mathcal{R}(\mathcal{V}_L) \to \mathcal{R}(\mathcal{V}_H)$, define $\omega_\tau(\vec{X} \leftarrow \vec{x}) = \vec{Y} \leftarrow \vec{y}$ if there exists $\vec{Y} \subseteq \mathcal{V}_H$ and $\vec{y} \in \mathcal{R}(\vec{Y})$ such that $\tau(Rst(\mathcal{V}_L, \vec{x})) = Rst(\mathcal{V}_H, \vec{y})$ (as usual, given $T \subseteq \mathcal{R}(\mathcal{V}_L)$, we define $\tau(T) = \{\tau(\vec{v}_L) : \vec{v}_L \in T\}$). It is easy to see that, given $\vec{X}$ and $\vec{x}$, there can be at most one such $\vec{Y}$ and $\vec{y}$. If there does not exist such a $\vec{Y}$ and $\vec{y}$, we take $\omega_\tau(\vec{X} \leftarrow \vec{x})$ to be undefined. Let $\mathcal{I}_L^\tau$ be the set of interventions for which ω_τ is defined, and let $\mathcal{I}_H^\tau = \omega_\tau(\mathcal{I}_L^\tau)$. ∎

It is straightforward to check that in Example 3.2, ω_τ is defined on interventions to A_1, A_2, and on these interventions it is the identity (and thus agrees with ω as defined in that example), but it is also defined on simultaneous interventions on $X_1 - X_{33}$, $X_{34} - X_{66}$, and $X_{67} - X_{99}$, and on T (as well as combinations of these interventions). In Example 3.3, the interventions on which ω_τ is defined are precisely those in the set $\mathcal{I}_L$ of that example; on these interventions, $\omega_\tau = \omega$.

Note that if τ is surjective, then it follows that $\omega_\tau(\emptyset) = \emptyset$, and for all $\vec{v}_L \in \mathcal{R}(\mathcal{V}_L)$, $\omega_\tau(\mathcal{V}_L \leftarrow \vec{v}_L) = \mathcal{V}_H \leftarrow \tau(\vec{v}_L)$.

Definition 3.13: $(M_H, \mathcal{I}_H)$ is a *τ-abstraction of* $(M_L, \mathcal{I}_L)$ if the following conditions hold:

- τ is surjective;
- there is a surjective function $\tau_{\mathcal{U}} : \mathcal{R}(\mathcal{U}_L) \to \mathcal{R}(\mathcal{U}_H)$ compatible with τ;
- $\mathcal{I}_H = \omega_\tau(\mathcal{I}_L)$. ∎

As intended, Examples 3.10 and 3.11 are not τ-abstractions; on the other hand, in Examples 3.2 and 3.3, M_H is a τ-abstraction of M_L.

Unlike exact and uniform transformations, τ-abstraction is a relation between causal models: the mapping ω is determined by τ, and there is no need to specify a probability distribution.

We can strengthen the notion of τ-abstraction to define a relation on basic causal models, by considering the largest possible sets of allowed interventions.

Definition 3.14: If M_H and M_L are basic causal models, then M_H is a *strong τ-abstraction* of M_L if $\mathcal{I}_H^\tau = \mathcal{I}_H^*$, the set of all high-level interventions, and $(M_H, \mathcal{I}_H^\tau)$ is a τ-abstraction of $(M_L, \mathcal{I}_L^\tau)$. ∎

The notion of strong τ-abstraction provides a clean, powerful relation between basic causal models. However, there are applications where the two additional requirements that make an abstraction strong are too much to ask. In the following example, neither requirement is satisfied.

Example 3.15: Consider an object in the earth's gravitational field. On the low level (M_L), there are three endogenous variables: V (velocity), H (height), and M (mass), and three corresponding exogenous variables, U_V, U_H, and U_M. The equations in M_L are $V = U_V$, $H = U_H$, and $M = U_M$. The high level captures the object's current energy. M_H contains endogenous variables K (kinetic energy) and P (potential energy), and two corresponding exogenous

variables, U_K and U_P. The equations in M_H are $K = U_K$ and $P = U_P$. We define $\tau : \mathcal{R}(\mathcal{V}_L) \to \mathcal{R}(\mathcal{V}_H)$ using the standard equations for kinetic energy and gravitational potential energy, so $\tau(v, h, m) = (\frac{1}{2}mv^2, 9.81mh)$. It is easy to see that τ is a surjection onto $\mathcal{R}(\mathcal{V}_H)$. We claim that M_H is not a strong τ-abstraction of M_L. To see why, consider interventions of the form $M \leftarrow m$ for $m > 0$. Applying Definition 3.12, we get that $\omega_\tau(M \leftarrow m) = \emptyset$, since $\tau(Rst(\mathcal{V}_L, m)) = \mathcal{V}_H$; by choosing v and h appropriately, we can still get all values in $\mathcal{V}_H$, as long as $m > 0$. We also clearly have that ω_τ maps the empty intervention in M_L to the empty intervention in M_H. With this, we can already show that $(M_H, \mathcal{I}_H^\tau)$ is not a uniform (τ, ω_τ)- transformation of $(M_L, \mathcal{I}_L^\tau)$. Suppose that $\Pr_L$ is a probability on $\mathcal{U}_L$ that puts probability 1 on $(1, 1, 1)$. For condition (1) in Definition 3.1 to hold for the intervention $M \leftarrow m$, the probability $\Pr_H$ on $\mathcal{U}_H$ must put probability 1 on $(.5m, 9.81m)$. But (1) must hold for all choices of m. This is clearly impossible.

Although M_H is not a strong τ-abstraction of M_L, we can easily construct a sensible and useful τ-abstraction between these models by simply not allowing interventions of the form $M \leftarrow m$ in the low-level model. Concretely, if we define $\mathcal{I}_L$ as containing the empty intervention and all interventions of the form $(V, H, M) \leftarrow (v, h, m)$, then ω_τ maps this to the set $\mathcal{I}_H$ that contains the empty intervention and all interventions of the form $(K, P) \leftarrow (k, p)$. ∎

As the following example shows, there also exist interesting cases where only the first requirement of Definition 3.14 is not satisfied. Roughly speaking, this is because some high-level variables are not logically independent, so not all high-level interventions are meaningful.

Example 3.16: Suppose that we have a 100×100 grid of pixels, each of which can be black or white. In the low-level model, we have 10,000 endogenous variables X_{ij}, for $1 \leq i, j \leq 100$, and 10,000 corresponding exogenous variables U_{ij} for $1 \leq i, j \leq 100$, with the obvious equations $X_{ij} = U_{ij}$. We would expect there to be other variables that are affected by the X_{ij}s (e.g., what a viewer perceives), but for ease of exposition, we ignore these other variables in this example and focus only on the X_{ij} variables. Suppose that all we care about is how many of the pixels in the upper half of the grid are black and how many pixels in the left half of the grid are black. Thus, in the high-level model, we have variables UH and LH whose range is $\{0, \ldots, 5000\}$. Because of the dependencies between UH and LH, there is a single exogenous variable that determines their values, which are pairs (m, m') such that $0 \leq m, m' \leq 5,000$ and $|m - m'| \leq 2,500$. Now we have an obvious map τ from low-level states to high-level states. We claim that $(M_H, \mathcal{I}_H^\tau)$ is a τ-abstraction of $(M_L, \mathcal{I}_L^\tau)$, where $\mathcal{I}_L^\tau$ consists of the empty intervention and interventions that simultaneously set all the variables in the upper half and left half (i.e., all variables X_{ij} with $1 \leq i \leq 50$ or $51 \leq j \leq 100$) and an arbitrary subset of the variables in the bottom right. Given a nonempty intervention $\vec{X} \leftarrow \vec{x}$ of this form, $\omega_\tau(\vec{X} \leftarrow \vec{x}) = (UH \leftarrow m, LH \leftarrow m')$, where m is the number of X_{ij} variables set to 1 with $1 \leq i \leq 50$ and m' is the number of X_{ij} variables set to 1 with $51 \leq j \leq 100$;

how the variables in the bottom right are set in $\vec{X} \leftarrow \vec{x}$ is irrelevant. Thus, $\mathcal{I}_H$ consists of interventions of the form $(UH \leftarrow m, LH \leftarrow m')$, where $1 \leq m, m' \leq 5000$ and $|m - m'| \leq 2500$. It is straightforward to check that there is no low-level intervention $\vec{X} \leftarrow \vec{x}$ such that $\omega_\tau(\vec{X} \leftarrow \vec{x}) = UH \leftarrow m$. For suppose that $\omega_\tau(\vec{X} \leftarrow \vec{x}) = UH \leftarrow m$. Then $\tau(Rst(\mathcal{V}_L, \vec{x})) = \{(m, m') : 1 \leq m' \leq 5000\}$. This means that $(m, m') \in \tau(Rst(\mathcal{V}_L, \vec{x}))$ for some m' such that $|m - m'| > 2500$, which is a contradiction. A similar argument shows that no intervention of the form $LH \leftarrow m'$ can be in $\mathcal{I}_H^\tau$. It is straightforward to check that $(M_H, \mathcal{I}_H^\tau)$ is a uniform (τ, ω_τ)-transformation of $(M_L, \mathcal{I}_L^\tau)$, so $(M_H, \mathcal{I}_H^\tau)$ is a τ-abstraction of $(M_L, \mathcal{I}_L^\tau)$, however it is clearly not a strong τ-abstraction of M_L. ∎

The problem here is that although M_H has variables UH and LH, we can only intervene on them simultaneously. It may make sense to consider such interventions if we want a visual effect that depends on both the number of black pixels in the upper half and the number of black pixels in the left half. But it is worth noting that if we consider a high-level model M_H' with only a single variable ULH that counts the number of pixels that are black in the upper half and the left half altogether, then M_H' is a strong τ-abstraction of M_L with the obvious map τ.

The full paper (posted on arxiv) gives an example where the second requirement of Definition 3.14 is not satisfied.

3.4 From micro-variables to macro-variables

Roughly speaking, the intuition for clustering micro-variables into macro-variables is that in the high-level model, one variable captures the effect of a number of variables in the low-level model. This makes sense only if the low-level variables that are being clustered together "work the same way" as far as the allowable interventions go. The following definition makes this precise.

Definition 3.17 : M_H is a *constructive* τ-abstraction of M_L if M_H is a strong τ-abstraction of M_L and, if $\mathcal{V}_H = \{Y_1, \ldots, Y_n\}$, then there exists a partition $P = \{\vec{Z}_1, \ldots, \vec{Z}_{n+1}\}$ of $\mathcal{V}_L$, where $\vec{Z}_1, \ldots, \vec{Z}_n$ are nonempty, and mappings $\tau_i : \mathcal{R}(\vec{Z}_i) \to \mathcal{R}(Y_i)$ for $i = 1, \ldots, n$ such that $\tau = (\tau_1, \ldots, \tau_n)$; that is, $\tau(\vec{v}_L) = \tau_1(\vec{z}_1) \cdot \ldots \cdot \tau_n(\vec{z}_n)$, where $\vec{z}_i$ is the projection of $\vec{v}_L$ onto the variables in $\vec{Z}_i$, and $\cdot$ is the concatenation operator on sequences. M_H is a *constructive abstraction* of M_L if it is a *constructive τ-abstraction* of M_L for some τ. ∎

In this definition, we can think of each $\vec{Z}_i$ as describing a set of microvariables that are mapped to a single macrovariable Y_i. The variables in $\vec{Z}_{n+1}$ (which might be empty) are ones that are marginalized away.

By definition, every constructive τ-abstraction is a strong τ-abstraction. We conjecture that a converse to this also holds: that is, if M_H is a strong τ-abstraction of M_L, that perhaps satisfies a few minor technical conditions, then it will in fact be a constructive τ-abstraction of M_L. However, we have not proved this result yet.

We suspect that constructive τ-abstractions are the notion of abstraction that will arise most often in practice. All three

of the examples discussed by RW$^+$ (one of which is Example 3.3) are constructive abstractions. We can easily extend Example 3.2 by adding low-level and high-level interventions to make it a constructive abstraction as well.

4 Discussion and Conclusions

We believe that getting a good notion of abstraction will be critical in allowing modelers to think at a high level while still being faithful to a more detailed model. As the analysis of this paper shows, there are different notions of abstraction, that relate causal models at different levels of detail. For example, τ-abstraction is a relation between basic causal models, while a uniform (τ, ω)- transformation relates causal models, and RW$^+$'s notion of exact transformation relates probabilistic causal models. Although our final notion of constructive abstraction is the cleanest and arguably easiest to use, we believe that there exist applications for which the weaker abstraction relations are more appropriate. More work needs to be done to understand which abstraction relation is most suitable for a given application. We hope that the definitions proposed here will help clarify the relevant issues. They should also shed light on some of the recent discussions of higher-level causation in communities ranging from physics to philosophy (see, e.g., (Fenton-Glynn 2017; Hoel, Albantakis, and Tononi 2013).

In fact, we see the current paper as laying the formal groundwork for several interesting topics that we intend to explore in future work. First, we hope to generalize the abstraction relation to a notion of *approximate abstraction*, given that in most real-life settings the mappings between different levels are only approximately correct. Second, our framework makes it possible to explore whether the notion of actual causation could be applied across causal models, rather than merely within a single causal model. For example, it seems to be useful to think of an event in a low-level model as causing an event in a high-level model. Third, abstracting causal models of large complexity into simpler causal models with only a few variables is of direct relevance to the increasing demand for explainable AI, for in many situations the problem lies not in the fact that no causal model is available, but in the fact that the only available model is too complicated for humans to understand.

A Appendix: Proofs

Proof of Lemma 2.4: Let $M = ((\mathcal{U}, \mathcal{V}, \mathcal{R}), \mathcal{F}, \mathcal{I}, \mathrm{Pr})$ and define $M' = ((\mathcal{U}', \mathcal{V}, \mathcal{R}'), \mathcal{F}', \mathcal{I}, \mathrm{Pr}')$ as follows. $\mathcal{U}'$ has one exogenous variable for each endogenous variable in $\mathcal{V}$. Taking $\mathcal{V} = \{Y_1, \ldots, Y_n\}$, we take U_i to be the exogenous

variable corresponding to Y_i. Let $\mathcal{U}' = \{U'_1, \ldots, U'_n\}$. We take $\mathcal{R}(U'_i) = \mathcal{R}(\mathcal{U})$ for $i = 1, \ldots, n$ (so the set of possible values for each variable U'_i is the set of all contexts in M). If $\vec{z} \in \mathcal{R}(\mathcal{V} - \{Y_i\})$, we define $F'_{Y_i}(\vec{z}, \vec{u}_1, \ldots, \vec{u}_n) = F_{Y_i}(\vec{z}, \vec{u}_i)$. (Note that here $\vec{u}_i \in \mathcal{R}(\mathcal{U}) = \mathcal{R}(U'_i)$.) Thus, it is clear that the only exogenous variable that the value of Y_i in M' depends on is U'_i, so M' has uev, as desired. $\Pr'$ places probability 0 on a context $(\vec{u}_1, \ldots, \vec{u}_n)$ unless $\vec{u}_1 = \ldots = \vec{u}_n$, and $\Pr'(\vec{u}, \ldots, \vec{u}) = \Pr(\vec{u})$. It is almost immediate that, with these choices, $M \sim M'$. ∎

Proof of Proposition 3.6: Suppose that M_H is a uniform $(\tau\text{-}\omega)$-transformation of M_L. Say that $\vec{u}_L$ and $\vec{u}_H$ *correspond* if $\tau(M_L(\vec{u}_L, \vec{Y} \leftarrow \vec{y})) = M_H(\vec{u}_H, \omega(\vec{Y} \leftarrow \vec{y}))$ for all interventions $\vec{Y} \leftarrow \vec{y} \in \mathcal{I}_L$.

We claim that for all $\vec{u}_L \in \mathcal{R}(\mathcal{U}_L)$, there exists at least one $\vec{u}_H \in \mathcal{R}(\mathcal{U}_H)$ that corresponds to $\vec{u}_L$. To see this, fix $\vec{u}_L \in \mathcal{R}(\mathcal{U}_L)$. Let $\Pr_L$ give $\vec{u}_L$ probability 1. Then for each intervention $\vec{Y} \leftarrow \vec{y} \in \mathcal{I}_L$, the distribution $\Pr_L^{\vec{Y} \leftarrow \vec{y}}$ gives probability 1 to $M_L(\vec{u}_L, \vec{Y} \leftarrow \vec{y})$. Let $\Pr_H$ be a probability distribution such that $(M_H, \Pr_H)$ is an exact $(\tau\text{-}\omega)$-transformation of $(M_L, \Pr_L)$. Since $\tau(\Pr_L^{\vec{Y} \leftarrow \vec{y}}) = \Pr_H^{\omega(\vec{Y} \leftarrow \vec{y})}$, it follows that $\Pr_H^{\omega(\vec{Y} \leftarrow \vec{y})}$ gives probability 1 to $\tau(M_L(\vec{u}_L, \vec{Y} \leftarrow \vec{y}))$, and hence also to the set $\mathcal{U}_H^{\vec{u}_L, \vec{Y} \leftarrow \vec{y}} = \{\vec{u}_H : M_H(\vec{u}_H, \omega(\vec{Y} \leftarrow \vec{y})) = \tau(M_L(\vec{u}_L, \vec{Y} \leftarrow \vec{y}))\}$. Since there are only finitely many interventions in $\mathcal{I}_L$, $\cap_{\vec{Y} \leftarrow \vec{y} \in \mathcal{I}_L} \mathcal{U}_H^{\vec{u}_L, \vec{Y} \leftarrow \vec{y}}$ also has probability 1, and thus must be nonempty. Choose $\vec{u}_H \in \cap_{\vec{Y} \leftarrow \vec{y} \in \mathcal{I}_L} \mathcal{U}_H^{\vec{Y} \leftarrow \vec{y}}$. By construction, $\vec{u}_H$ corresponds to $\vec{u}_L$.

Define $\tau_\mathcal{U}$ by taking $\tau_\mathcal{U}(\vec{u}_L) = \vec{u}_H$, where $\vec{u}_H$ corresponds to $\vec{u}_L$. (If more than one tuple $\vec{u}_H$ corresponds to $\vec{u}_L$, then one is chosen arbitrarily.) It is now straightforward to check that $(M_H, \tau_\mathcal{U}(\Pr_L))$ is an exact $(\tau\text{-}\omega)$-transformation of $(M_L, \Pr_L)$. We leave details to the reader. ∎

Proof of Theorem 3.8: To show that (a) implies (b), suppose that M_H is a uniform $(\tau\text{-}\omega)$-transformation of M_L. Let $\tau_\mathcal{U} : \mathcal{R}(\mathcal{U}_L) \to \mathcal{R}(\mathcal{U}_H)$ be the function guaranteed to exist by Proposition 3.6. We must show that for all $\vec{Y} \leftarrow \vec{y} \in \mathcal{I}_L$, $\vec{u}_L \in \mathcal{R}(\mathcal{U}_L)$,

$$\tau(M_L(\vec{u}_L, \vec{Y} \leftarrow \vec{y})) = M_H(\tau_\mathcal{U}(\vec{u}_L), \omega(\vec{Y} \leftarrow \vec{y})).$$

Fix $\vec{u}_L \in \mathcal{R}(\mathcal{U}_L)$. From the construction of $\tau_\mathcal{U}$ in the proof of Proposition 3.6, it follows that $\vec{u}_L$ and $\tau_\mathcal{U}(\vec{u}_L)$ correspond, which, by definition, means that $\tau(M_L(\vec{u}_L, \vec{Y} \leftarrow \vec{y})) = M_H(\tau_\mathcal{U}(\vec{u}_L), \omega(\vec{Y} \leftarrow \vec{y}))$ for all interventions $\vec{Y} \leftarrow \vec{y} \in \mathcal{I}_L$.

To show that (b) implies (a), suppose that (b) holds. Given a distribution $\Pr_L$ on $\mathcal{R}(\mathcal{U}_L)$, let $\Pr_H = \tau_\mathcal{U}(\Pr_L)$. It suffices to show that $(M_H, \Pr_H)$ is an exact $(\tau\text{-}\omega)$-transformation of $(M_L, \Pr_L)$. Thus, we must show that for every intervention $\vec{Y} \leftarrow \vec{y} \in \mathcal{I}_L$, we have $\Pr_H^{\omega(\vec{Y} \leftarrow \vec{y})} = \tau(\Pr_L^{\vec{Y} \leftarrow \vec{y}})$. Straight-

forward computations now show that

$$
\begin{aligned}
& \Pr_H^{\omega(\vec{Y} \leftarrow \vec{y})}(\vec{v}_H) \\
=& \Pr_H(\{\vec{u}_H : M_H(\vec{u}_H, \omega(\vec{Y} \leftarrow \vec{y})) = \vec{v}_H\}) \\
=& \Pr_L(\{\vec{u}_L : M_H(\tau_\mathcal{U}(\vec{u}_L), \omega(\vec{Y} \leftarrow \vec{y})) = \vec{v}_H\}) \\
=& \Pr_L(\{\vec{u}_L : \tau(M_L(\vec{u}_L, \vec{Y} \leftarrow \vec{y})) = \vec{v}_H\}) \\
=& \Pr_L(\{\vec{v}_L : \tau(\vec{v}_L) = \vec{v}_H \text{ and} \\
& \quad \exists \vec{u}_L(M_L(\vec{u}_L, \vec{Y} \leftarrow \vec{y}) = \vec{v}_L)\}) \\
=& \Pr_L^{\vec{Y} \leftarrow \vec{y}}(\{\vec{v}_L : \tau(\vec{v}_L) = \vec{v}_H\}) \\
=& \tau(\Pr_L^{\vec{Y} \leftarrow \vec{y}})(\vec{v}_H),
\end{aligned}
$$

as desired. ∎

References

Binahashemi, B.; de Giacomo, G.; and Lespérance, Y. 2017. Abstraction in situation calculus action theories. In *Proc. Thirty-First National Conference on Artificial Intelligence (AAAI '17)*, 1048–1055.

Chalupka, K.; Eberhardt, F.; and Perona, P. 2015. Visual causal feature learning. In *Proc. 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2015)*, 181–190.

Chalupka, K.; Eberhardt, F.; and Perona, P. 2016. Multi-level cause-effect systems. In *Proc. 19th Int. Conf. on Artificial Intelligence and Statistics (AISTATS 2016)*, 361–369.

Fenton-Glynn, L. 2017. Is there high-level causation? *Ergo* 4(30):845–898.

Halpern, J. Y., and Pearl, J. 2005. Causes and explanations: a structural-model approach. Part I: Causes. *British Journal for Philosophy of Science* 56(4):843–887.

Halpern, J. Y. 2016. *Actual Causality*. Cambridge, MA: MIT Press.

Hoel, E. P.; Albantakis, L.; and Tononi, G. 2013. Quantifying causal emergence shows that macro can beat micro. *Proc. National Academy of Science* 110(49):19790–19795.

Iwasaki, Y., and Simon, H. A. 1994. Causality and model abstraction. *Artificial Intelligence* 67(1):143–194.

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Rubenstein, P. K.; Weichwald, S.; Bongers, S.; Mooij, J. M.; Janzing, D.; Grosse-Wentrup, M.; and Schölkopf, B. 2017. Causal consistency of structural equation models. In *Proc. 33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*.