A formal framework for representing mechanisms?[†‡]

Alexander Gebharter

Düsseldorf Center for Logic and Philosophy of Science (DCLPS)

Abstract: In this paper I tackle the question of how mechanisms can be represented within a causal graph framework. I begin with a few words on mechanisms and some of their characteristic properties. I then concentrate on how one of these characteristic properties, *viz.* the hierarchic order of mechanisms (mechanisms frequently consist of several submechanisms), can be represented within a causal graph framework. I illustrate an answer to this question proposed by Casini, Illari, Russo, & Williamson (2011) and demonstrate on an example that their formalism, though nicely capturing the hierarchic order of mechanisms, does not support two important features of nested mechanisms: (i) a mechanism's submechanisms are typically causally interacting with other parts of said mechanism, and (ii) intervening in some of a mechanism's parts should have some influence on the phenomena the mechanism as a whole brings about. Finally, I sketch an alternative approach capable of taking properties (i) and (ii) into account and demonstrate this on the above-mentioned exemplary mechanism.

1. Introduction

In many scientific fields phenomena are explained and/or predicted by pointing at their underlying mechanisms. Such mechanisms are thought of as concrete entities located at specific regions in space-time which produce the respective phenomena. They are characterized by formulations like the following:[1]

> A mechanism underlying a behavior is a complex system which produces that behavior by of

[†] This is a draft paper. The final version of this paper is published under the following bibliographical data: Gebharter, A. (2014). A formal framework for representing mechanisms? *Philosophy of Science, 81*(1), 138-153. doi:10.1086/674206.

[‡] An earlier version of this paper won a best paper award at the *8th International Conference of the Association for Analytic Philosophy* (GAP.8).

[1] For alternative formulations see, for example, Bechtel & Abrahamsen (2005), Illari & Williamson (2012), and Machamer, Darden, & Craver (2000).

the interaction of a number of parts according to direct causal laws. (Glennan, 1996, p. 52)

According to mechanists, mechanisms are dynamic causal systems; they are wholes consisting of several spacio-temporally arranged and interacting parts producing certain behavior. Beside these properties, mechanisms are oftentimes (but not always) self-regulating systems including a lot of feedback loops. Typically (but not necessarily), they are also hierarchically organized (i.e., they consist of several interacting submechanisms which may themselves be built up of submechanisms etc.). The more is known about the structure of these submechanisms, the more accurate the predictions of the phenomena these mechanisms bring about will typically be.

Although characterizations, like the one formulated by Glennan (1996) above, are intuitively quite clear, they are not as helpful as one may hope for when it comes to quantitatively precise explanation/prediction of phenomena of interest. This deficit can easily be seen by means of the following example: The question of why a car speeds up when the gas pedal is pressed can be answered by pointing at/describing the underlying mechanism (i.e., the motor and how it is connected to the gas pedal, the wheels, the gas tank, etc.), but questions including numerical details like why the acceleration of the car is *a* when the gas pedal is pressed with pressure *p* cannot be answered that easily. The answer to a question like the latter requires a formalism capable of capturing/computing the numerical details/effects of specific manipulations of said mechanism.

Such a formalism must be able to represent the above-mentioned characteristic properties of mechanisms in an adequate way. In a (2011) paper Casini, Illari, Russo, & Williamson propose to model mechanisms on the basis of so-called recursive Bayesian networks, which were originally developed by Williamson & Gabbay (2005) to model nested causal relationships. In doing so they focus on an adequate representation of the hierarchic structure of mechanisms and represent submechanisms by means of a recursive Bayesian network's vertices. I will briefly introduce the formal preliminaries needed to take a closer look at their approach and explain their account on a very simple exemplary toy mechanism in sec. 2. In sec. 3 I will highlight two problems with Casini *et al.*'s approach: Their approach does (i) not allow for a graphical representation of how a mechanism's macro variables are causally connected to the mechanism's causal micro structure, which is essential when it comes to mechanistic explanation, and it (ii) leads to the fatal consequence that a mechanism's macro variables' values cannot be changed by any intervention on the mechanism's micro structure whatsoever, and thus, contradicts the fact that scientists regularly perform so-

called bottom-up experiments to investigate which are the mechanism's constitutively relevant parts. In sec. 4 I present an alternative approach for modeling nested mechanisms: Submechanisms should not be represented by means of a causal graph's vertices, like in Casini *et al.*'s approach, but rather by means of a causal graph's edges. I finally demonstrate on the above-mentioned exemplary mechanism that this approach does not fall prey to problems (i) and (ii).

2. Bayesian networks, recursive Bayesian networks, and the RBN approach

A *Bayesian network* (BN) is a triple $\langle V,E,P\rangle$ that satisfies the so-called *Markov condition* (MC). $G = \langle V,E\rangle$ is a *graph* whose *vertices* (i.e., the elements of V) are random variables that may take a number of different values, while E is a binary relation on V ($E \subseteq V \times V$). E's elements $\langle X,Y\rangle$ are called *edges* and can be graphically represented via different kinds of lines and/or arrows in G. A BN's associated graph is always a *directed acyclic graph* (DAG), i.e., a graph whose edges are arrows ($X \rightarrow Y$) and that does not contain a substructure of the form $X \rightarrow ... \rightarrow X$. P is a joint probability distribution over the random variables in V.

(1) *Markov condition:* $\langle V,E,P\rangle$ satisfies MC if and only if $INDEP(X, V - Des(X) | Par(X))$ holds for all $X \in V$. (Spirtes, Glymour, & Scheines 2000, p. 11)

'$Des(X)$' stands for the descendants (i.e., the successors) of X in graph $G = \langle V,E\rangle$, '$Par(X)$' for the parents (i.e., the direct predecessors) of X in graph $G = \langle V,E\rangle$, and '$INDEP(X,Y|Z)$' for probabilistic independence of X and Y conditional on Z (i.e., $P(x|y,z) = P(x|z)$ for all X-, Y-, and Z-values x, y, and z, respectively, provided $P(y,z) > 0$). BNs can be causally interpreted, i.e., they can be understood as a certain type of causal model. When doing so, a BN's associated graph G represents the system of interest's causal structure. '$X \rightarrow Y$' in such a *causal graph* G stands for 'X is a *direct cause* of Y in G', and a chain of (one or more) arrows (i.e., a *directed path*) going from X to Y for 'X is a (direct/indirect) *cause* of Y in G'. A structure of the form $X \leftarrow ... \leftarrow Z \rightarrow \ldots \rightarrow Y$ is called a *common cause path* between X and Y.

When one uses BNs for causal modeling, also MC is causally interpreted. Under its causal interpretation MC becomes the so-called *causal Markov condition* (CMC) that is satisfied by a causal model $\langle V,E,P\rangle$ if and

only if every $X \in V$ is probabilistically independent of all its non-effects conditional on its direct causes (cf. Spirtes *et al.*, 2000, p. 29).[2]

The graph $G = \langle V,E \rangle$ of a BN satisfying MC/CMC determines the following Markov factorization:[3]

(2) $P(x_1,\ldots,x_n) = \prod_i P(x_i|par(X_i))$

A *recursive Bayesian network* (RBN) is a BN in which the values of variables in V can be BNs themselves. Such variables are called *network variables*, while variables that do not have BNs as values are called *simple variables*. Casini *et al*. (2011) suggest to represent a mechanism by an RBN $\langle V,E,P \rangle$ and a submechanism by a network variable $X \in V$ whose values are BNs representing the possible states of this submechanism. They propose, in addition to the causal interpretation of MC, an additional modeling assumption, the *recursive causal Markov condition* (RCMC):

(3) *Recursive causal Markov condition:* $\langle V,E,P \rangle$ satisfies RCMC if and only if $INDEP(X,NID(X)|DSup(X) \cup Par(X))$ holds for all $X \in V$. (Casini *et al*., 2011, p. 11)

$NID(X)$ is the set of *non-inferiors-or-descendants* of X, i.e., the set of random variables that are neither inferiors nor descendants of X. The *inferiors* of X are the variables of a lower-level BN representing states of the submechanism described by X at the higher level, the variables of the lower-level BNs representing states of submechanisms of this submechanism, etc. $DSup(X)$ is the set of *direct superiors* of X. $DSup(X)$ contains those variables of the next level up BN representing a submechanism whose states are described by lower-level BNs including X. (For an illustration of these notions see the water dispenser example introduced below.) Casini *et al.* (2011, sec. 4) suggest to interpret the inferiority/superiority relation as constitutive relevance in the sense of Craver (2007a, 2007b).

Let me now briefly explain how probabilistic interlevel explanation/prediction works in Casini *et al.*'s (2011) RBN approach. One therefore needs to define $\boldsymbol{V} = \{X_1,\ldots,X_m\}$ as the RBN $\langle V,E,P \rangle$'s variable set V

[2] CMC is the generalization of an idea that can be traced back to Reichenbach's (1956) book *The Direction of Time*: Correlated effects are screened off each other by conditionalizing on their common causes; effects are screened off their indirect causes by conditionalizing on their direct causes.

[3] '$par(X_i)$' stands for the instantiation of X_i's parents to their values $x_1,\ldots,x_n$ on the left hand side of the equation.

under the transitive closure of the inferiority relation.[4] Let $\boldsymbol{N} = \{X_{j1},\ldots,X_{jk}\}$ be the set of network variables in $\boldsymbol{V}$. Then for every instantiation $\boldsymbol{n} = x_{j1},\ldots,x_{jk}$ of network variables in $\boldsymbol{N}$ a simple BN can be constructed: the *flattening* of the RBN w.r.t. $\boldsymbol{n}$ ($\boldsymbol{n}\downarrow$). The nodes of this new BN $\boldsymbol{n}\downarrow$ are the simple variables in $\boldsymbol{V}$ together with the instantiations $\boldsymbol{n} = x_{j1},\ldots,x_{jk}$ of the network variables in $\boldsymbol{N}$. $\boldsymbol{n}\downarrow$'s set of edges contains an arrow pointing from X to Y if and only if X is a parent or direct superior of Y in the RBN. $\boldsymbol{n}\downarrow$'s probability distribution is determined by the following equation:

(4) $P(x_i|par(X_i),dsup(X_i)) = P_{xjl}(x_i|par(X_i))$, where X_{jl} are the direct superiors of X_i.

The flattenings $\boldsymbol{n}\downarrow$ of an RBN determine a unique probability distribution over $\boldsymbol{V} = \{X_1,\ldots,X_m\}$ that allows for quantitative reasoning across the diverse levels of the mechanism represented by the RBN:[5]

(5) $P(x_1,\ldots,x_m) = \prod_i P(x_i|par(X_i),dsup(X_i))$

Let me now briefly illustrate how the modeling approach proposed by Casini *et al.* (2011) works on a very simple toy example, *viz.* the water dispenser mechanism. This device normally dispenses cold water and water close to the room temperature when its tempering button is pressed. The water dispenser can be represented by an RBN whose top level graph is depicted in figure 1.

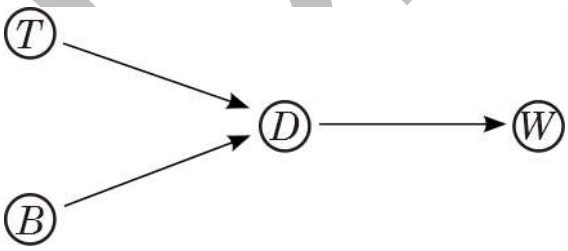

Figure 1

Figure2

T represents the room temperature, *B* = 1/0 stands for whether the tempering button is pressed or not, and

[4] The *transitive closure* R^* of a binary relation R can be defined as $R^* = \{\langle u,v\rangle: \exists w_1,...,\exists w_n(\langle u,w_1\rangle \in R \wedge \ldots \wedge \langle w_n,v\rangle \in R)\}$.
[5] The probabilities $P(x_i|par(X_i),dsup(X_i))$ on the right hand side of this equation are determined by the flattening induced by $x_1,\ldots,x_m$.

W for the temperature of the water dispensed. *D* is a network variable that represents a submechanism, *viz.* the water dispenser's water temperature regulation unit. This regulation unit consists of two lower-level parts: a temperature sensor (*S*) and a heater (*H*). *D* has two possible values: BN_1 (water temperature is regulated and thus, one gets water close to the room temperature) and BN_0 (water temperature is not regulated and cold water is dispensed as a result). BN_1 and BN_0 are two BNs with the same topological structure (depicted in figure 2), but with different associated probability distributions. If $D = BN_1$, then the heater is working on a level corresponding to the input of the temperature sensor. If $D = BN_0$, then *H* is probabilistically insensitive to *S*. Note that the singleton of *D* is the set of direct superiors of *S* and *H* ($\{D\} = DSup(S) = DSup(H)$) in our exemplary mechanism, while $\{S,H\}$ is the set of inferiors of *D* ($\{S,H\} = Inf(D)$).

When one wants to use the RBN approach for probabilistic predictions across the levels of a mechanism, one first has to construct the RBN's flattenings as described above. Figure 3 shows the flattening of the RBN w.r.t. $D = BN_1$. Note that the two interlevel arrows from *D* to *S* and from *D* to *H* stand for the direct superiority/inferiority relation[6] and should not be causally interpreted: *S* and *H* are not effects of *D*, they rather stand for constitutively relevant parts of the submechanism represented by *D*. To indicate this fact, the arrows are dashed in figure 3.

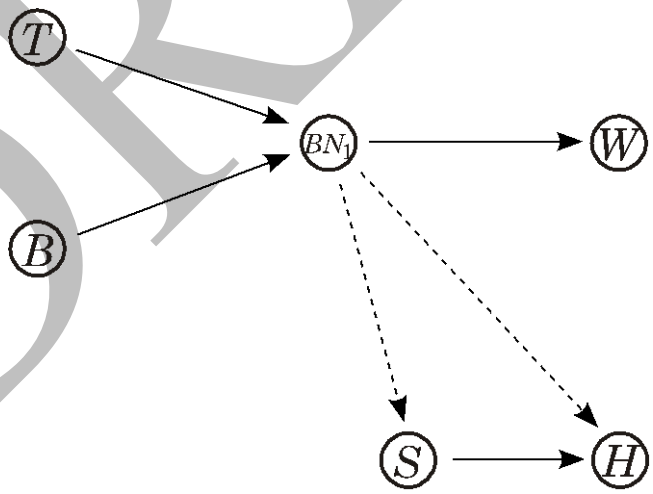

Figure 3

According to (4), the conditional probability distribution of this flattening is $P(T)$, $P(B)$, $P(D = BN_1) = 1$,

[6] If such an interlevel arrow is pointing from a variable *X* to a variable *Y*, then *X* is a direct superior of *Y* and *Y* is a direct inferior of *X* in the RBN. If a directed path of such interlevel arrows is going from *X* to *Y*, then *X* is a (direct or indirect) superior of *Y* and *Y* is a (direct or indirect) inferior of *X*.

$P(W|D = BN_1)$, $P(S) = P_{D = BN1}(S)$, $P(H) = P_{D = BN1}(H|S)$. The conditional probability distribution of the flattening of the RBN w.r.t. $D = BN_0$ is $P(T)$, $P(B)$, $P(D = BN_0) = 1$, $P(W|D = BN_0)$, $P(S) = P_{D = BN0}(S)$, $P(H) = P_{D = BN0}(H|S)$. According to (5), the two flattenings of the RBN determine a joint probability distribution over $\boldsymbol{V} = \{T,B,D,W,S,H\}$, *viz.* $P(T,B,D,W,S,H) = P(T) \cdot P(B) \cdot P(D|T,B) \cdot P(W|D) \cdot P(S|D) \cdot P(H|S,D)$, where the probabilities on the right hand side of the equation are determined by the flattening induced by *T,B,D,W,S,H*. This probability distribution can be used for quantitative prediction across the two levels of our exemplary mechanism.

3. Two problems with the RBN approach

Let me now expose the two deficits of the RBN approach announced in sec. 1. Problem (i): While RBNs clearly allow for quantitative reasoning across the diverse levels of mechanisms, they do not tell us how exactly submechanisms are causally connected to their mechanisms. In case of the water dispenser example, for instance, the RBN's graph does neither tell us how *T* and *B* causally influence *S* and *H*, nor how *S* and *H* are causally relevant for *W*, i.e., there are no arrows between those variables in the RBN's graph and it is unclear over which causal paths probabilistic influence from *T* and *B* is propagated through the mechanism's micro structure to *W*. But is the graphical representation of such causal information required at all? Is it not sufficient that the RBN captures the probabilistic dependencies between the variables in **V** = {*T,B,D,W,S,H*}? The answer to this latter question is a negative one. One of the reasons for this is simply that mechanistic explanation requires information about how exactly, i.e., over which causal pathways, certain inputs to the system influence the mechanism's micro structure and how changes in this micro structure bring about the phenomenon (or phenomena) of interest at the macro structure (cf., e.g., Bechtel, 2007, sec. 3).[7] In causal models this information is typically provided by the model's associated probability distribution *together* with its graph's topology. Illustrated on our example: If our RBN model adequately represents the water dispenser mechanism, then the information that the tempering button is not pressed ($B = 0$) will screen *W* off from *T*. (The room temperature is only relevant for the temperature of the water dispensed when the tempering button is pressed.) The RBN's associated probability distribution may give us

[7] There is an analogy in the discussion on scientific explanation: For explaining an event e_2 by referring to an earlier event e_1, knowing that e_1 increases e_2's probability is not enough; what one in addition has to know is that e_1 is *causally* relevant to e_2, one has to provide a model that shows *how* e_1 causes e_2 (cf. Salmon, 1984; Woodward, 2011, sec. 4).

the correct probabilistic dependencies/independencies, but its graph does not provide the causal information to mechanistically *explain* this probabilistic behavior. So the model does not tell us that the probabilistic influence of *T* on *W* breaks down because $B = 0$ fixes the value of *H* and *because H* lies on the only directed causal path from *T* to *W*.

The representation of such causal information in the model's graph is not only important for mechanistic explanation, but also when it comes to questions of manipulation and control. (Purely probabilistic models cannot distinguish between observation and manipulation; cf. Pearl 2009, sec. 1.3.1). So how, for example, could we intervene on the mechanism's micro structure in such a way that we can amplify or decrease certain external influences? If we want, for instance, to increase or decrease *T*'s causal effect on *W* in our exemplary mechanism, then the information (which is not captured by the RBN's graph) that *S* lies on a causal path from *T* to *W* is crucial. Such knowledge tells us that we can increase/decrease *T*'s effect on *W* by manipulating *S* in certain ways, e.g., by putting an additional heat source to the sensor *S* or by cooling *S*.[8]

Let me now illustrate problem (ii), which is presumably the more striking one of the two problems for the RBN approach: Recall that the probability distribution that allows for probabilistic reasoning across all levels of a mechanism is constructed via the flattenings of the RBN (see sec. 2). For our exemplary mechanism this probability distribution would be $P(T,B,D,W,S,H) = P(T) \cdot P(B) \cdot P(D|T,B) \cdot P(W|D) \cdot P(S|D) \cdot P(H|S,D)$, where the probabilities on the right hand side of the equation are determined by the flattening induced by *T,B,D,W,S,H*. This probability distribution can be captured by a BN with a graph like the one depicted in the box in figure 4. (Again, the continuous lines could, while the dashed ones should not be causally interpreted.) Now assume that one would, for example, intervene on *S* by means of an intervention variable I_S. Such an intervention on *S* would, and this can directly be read off the BN's associated graph's topology (depicted in figure 4), not have any probabilistic influence on any macro variable at all.

[8] Note that such amplification and/or decrease of a certain variable's influence on another one is not possible by means of so-called surgical or ideal interventions in the sense of Pearl (2009) or Woodward (2003); but it is by means of soft interventions (cf. Eberhardt & Scheines, 2007).

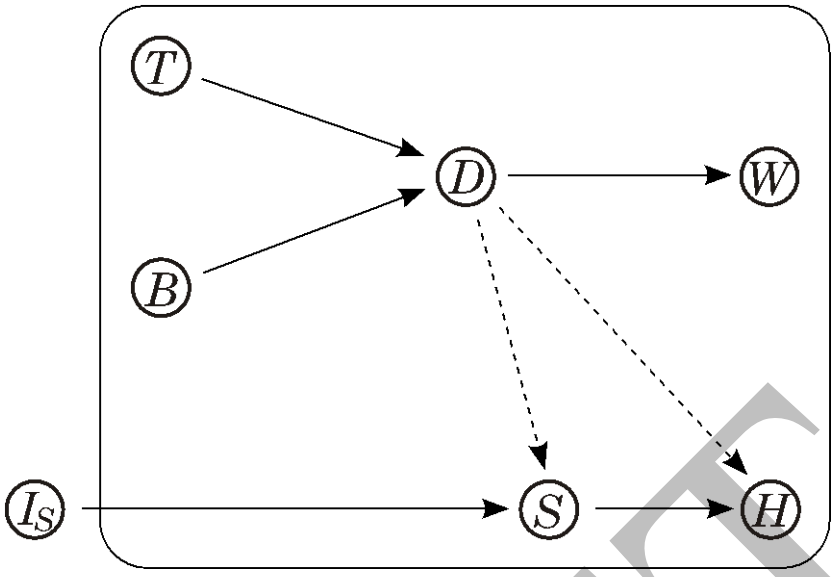

Figure 4

So, according to the RBN approach, intervening on a mechanism's micro variables does not have any probabilistic influence on any one of the macro variables whatsoever. This does not only contradict what we observe when looking at the (bottom-up) experiments scientists perform; it is also inconsistent with one of the core features of mechanisms: A mechanism's macro and its constitutively relevant micro behaviors should be mutually manipulable (cf. Craver 2007a, 2007b). Note that the inferiority relation is explicitly intended to represent constitutive relevance within the RBN approach (cf. Casini *et al.* 2011, sec. 4).

4. An alternative

Let me now propose an alternative to Casini *et al.*'s (2011) method for representing nested mechanisms. Instead of BNs I use causal models $\langle V,E,P\rangle$ whose graphs $G = \langle V,E\rangle$ are not restricted like the ones of BNs. In particular, the causal graphs $G = \langle V,E\rangle$ I use can contain two kinds of edges: $X \rightarrow Y$, which means that X is a direct cause of Y in the graph, and $X \leftrightarrow Y$, which means that X and Y are effects of a latent common cause, i.e., a cause of X and Y not represented within the graph's variable set V.[9] Contrary to Casini *et al.*, I suggest to represent mechanisms not by means of variables, but by means of *causal arrows*. So the simplest

[9] Note that the graph of a causal model that contains bidirected arrows does not anymore determine the Markov factorization (2). Causal models containing bidirected arrows will typically violate the Markov condition (1) as well as its causal interpretation, i.e., the causal Markov condition.

representation of a mechanism's top level would be a causal model $\langle V,E,P\rangle$ with graphical structure $X\rightarrow Y$ or $X\leftrightarrow Y$. In the first case, X would be the mechanism's *input*, Y its *output*, and the arrow '$\rightarrow$' would stand for the (not further specified) mechanism at work. In the latter case, X and Y would both represent different outputs produced by one or more not further specified (and maybe yet unknown) common causes. Also here, '$\leftrightarrow$' would stand for the mechanism at work.

To represent the mechanism's causal micro structure one can now assign a second causal model to the top level causal model $\langle V,E,P\rangle$ that specifies how exactly probabilistic influence between X and Y is propagated through the mechanism's causal micro structure. Both causal models must fit together with respect to the causal information contained in their associated graphs as well as with respect to the probabilistic information stored in their associated probability distributions. This is guaranteed by the following notion of a *restriction* of a causal model. This notion is basically a slightly modified version of Steel's (2005, p. 11) notion of a restricted graph complemented by conditions for bidirected arrows:

(6) $\langle V,E,P\rangle$ is a *restriction* of $\langle V^*,E^*,P^*\rangle$ if and only if

(a) $V \subset V^*$, and

(b) $P^*{\uparrow}V = P$,[10] and

(c) for all $X,Y \in V$:

(c.1) If there is a directed path from X to Y in $\langle V^*,E^*\rangle$ and no vertex on this path different from X and Y is in V, then $X\rightarrow Y$ in $\langle V,E\rangle$, and

(c.2) if X and Y are connected by a common cause path π in $\langle V^*,E^*\rangle$ or by a path π free of colliders[11] containing a bidirected edge in $\langle V^*,E^*\rangle$, and no vertex on this path π different from X and Y is in V, then $X\leftrightarrow Y$ in $\langle V,E\rangle$, and

(d) no path not implied by (c) is in $\langle V,E\rangle$.

Definition (6) determines for every causal model $\langle V^*,E^*,P^*\rangle$ and for every proper subset V of V^* a unique

[10] $P{\uparrow}V$ is the restriction of probability distribution P to variable set V.

[11] Z_l is called a *collider* on a causal path π if and only if π contains a subpath of the form $Z_k{*}\rightarrow Z_l \leftarrow {*}Z_m$, where the asterisk '*' is a meta-symbol standing for an arrowhead or an arrow's tail. So '$X{*}\rightarrow Y$', for example, stands for '$X\rightarrow Y$ or $X\leftrightarrow Y$'.

restriction $\langle V,E,P\rangle$. This restriction is called $\langle V^*,E^*,P^*\rangle$'s *restriction to V*. The introduced notion of a restriction allows for marginalizing out variables in such a way that the causal as well as the probabilistic information captured by the restricted model is preserved. $\langle V,E\rangle$ can be interpreted as a higher- and $\langle V^*,E^*\rangle$ as a lower-level mechanism's causal structure in (6). Condition (a) guarantees that the higher-level structure contains fewer variables than the lower-level one. (b) ensures that $\langle V,E,P\rangle$'s and $\langle V^*,E^*,P^*\rangle$'s probability distributions fit together, and (c) that also their associated causal structures do: Thanks to (c.1) all components of a mechanism represented at both levels are directly causally connected at the higher level whenever they are directly causally connected at the lower level; so no direct causal connection between two variables represented at both levels gets lost when going from the lower to the higher level. In addition it guarantees that there is a direct causal connection for every directed causal path in the lower-level structure whose intermediate components are not represented at the higher-level model's associated graph. (c.2) tells us when we have to draw a bidirected edge (↔) between two variables *X* and *Y* in the higher-level model's graph: Draw such a bidirected edge whenever there also is one at the lower level, if all variables on a common cause path of *X* and *Y* are marginalized out when going from the lower to the higher-level structure, or when all variables lying on a path at the lower level that indicates a latent common cause of *X* and *Y* are marginalized out.[12] (d) prevents causal connections at the higher level that do not have a counterpart at the lower level. Figure 5 illustrates how marginalizing out variables functions according to (6) by showing an exemplary causal structure and some of its possible restrictions.

Figure 5: According to (6), the graph of the restriction of a causal model with the graph depicted above would be $X\leftrightarrow Y\leftarrow Z\leftrightarrow W$ if one chooses to marginalize out *U*. It would be $X\leftrightarrow Y\leftarrow U\rightarrow W$ if marginalizing out *Z*, and $X\leftrightarrow Y\leftrightarrow W$ if marginalizing out *Z* and *U*. When one restricts the original model to $V = \{X,Z,U,W\}$, the resulting structure would be X $Z\leftarrow U\rightarrow W$ (without an edge between *X* and *Z*).

Let me now further develop the above mentioned idea of representing nested mechanisms by edges

[12] Here is an example of such a path: $Z_1\leftrightarrow Z_2$ in structure $X\leftarrow Z_1\leftrightarrow Z_2\rightarrow Y$ indicates a latent common cause of Z_1 and Z_2, and thus, also of *X* and *Y*. When marginalizing out Z_1 and Z_2, one has to draw a bidirected arrow between *X* and *Y* ($X\leftrightarrow Y$) to prevent this piece of causal information.

instead of vertices. For this purpose I introduce the following notion of a multi-level causal model (MLCM) that is based on the definition (6) of a restriction. I propose MLCMs as adequate means for representing the hierarchical organization of mechanisms. (Below I will demonstrate that MLCMs do not fall prey to the two problems of the RBN approach discussed in sec. 3):

(7) $\langle M_1 = \langle V_1,E_1,P_1\rangle,\ldots,M_n = \langle V_n,E_n,P_n\rangle\rangle$ is a *multi-level causal model* if and only if

(a) $M_1,\ldots,M_n$ are causal models, and

(b) every M_i with $1 < i \leq n$ is a restriction of M_1, and

(c) M_1 satisfies CMC.

According to (7), an MLCM is an *n*-tuple consisting of several causal models (condition (a)) which are intended to represent causal structures at different levels. According to (b), every causal model M_i in the ordering different from M_1 is a restriction of the first causal model M_1; so M_1 stands for the mechanism's lowest level, while every M_i different from M_1 represents one of its higher levels. Condition (c) captures a basic assumption of the causal nets approach, *viz.* that every robust probability distribution is produced (and, thus, can be explained) by some underlying causal model satisfying CMC (cf. Spirtes *et al.*, 2000, pp. 124f.). So an MLCM of a mechanism is complete only when all probabilistic dependencies and independencies of any higher-level model can be explained by a lowest level causal model M_1 that satisfies CMC.

(7) does not directly tell us much about the hierarchic organization of the mechanism and its submechanisms represented by the MLCM's causal models; it just tells us that M_1 stands for the lowest level. Fortunately, a unique *level graph* $G = \langle V,E\rangle$ can be constructed for every MLCM. Such a level graph is a kind of meta-graph that provides exactly the information requested above: information about the hierarchical relation of nested mechanisms represented by the MLCM:

(8) A graph $G = \langle V,E\rangle$ is called an MLCM $\langle M_1 = \langle V_1,E_1,P_1\rangle,\ldots,M_n = \langle V_n,E_n,P_n\rangle\rangle$'s *level graph* if and only if

(a) $V = \{M_1,\ldots,M_n\}$, and

(b) for all $M_i = \langle V_i,E_i,P_i\rangle$ and $M_j = \langle V_j,E_j,P_j\rangle$ in V: $M_i \rightarrow M_j$ in G if and only if $V_i \subset V_j$ and there

is no $M_k = \langle V_k,E_k,P_k\rangle$ in V such that $V_i \subset V_k \subset V_j$ holds.

According to (a), a level graph $G = \langle V,E\rangle$ is a graph over the causal models $M_1 = \langle V_1,E_1,P_1\rangle,\ldots,M_n = \langle V_n,E_n,P_n\rangle$ of an MLCM. (b) instructs one to draw a directed edge from one of these $M_i = \langle V_i,E_i,P_i\rangle$ to another $M_j = \langle V_j,E_j,P_j\rangle$ whenever V_i is a proper subset of V_j and there is no $M_k = \langle V_k,E_k,P_k\rangle$ in V such that V_k is a proper subset of V_j and a proper superset of V_i. So the directed paths in a level graph correspond to the set theoretical proper subset relation (i.e., $M_i \rightarrow \ldots \rightarrow M_j$ in G if and only if $V_i \subset V_j$). Because every causal model $M_i = \langle V_i,E_i,P_i\rangle$ of the MLCM different from $M_1 = \langle V_1,E_1,P_1\rangle$ is a restriction of M_1, the vertex set V_i of every such model M_i is a proper subset of V_1. So the level graph G will be a DAG containing only one vertex with no exiting arrows, *viz.* $M_1 = \langle V_1,E_1,P_1\rangle$, while there will be a directed path from every M_i different from M_1 to M_1.

Now some information about the hierarchical organization of causal models of an MLCM can be read off this MLCM's level graph G: Whenever there is a directed path from M_i to M_j in the level graph G, then M_i represents a higher-level causal structure than M_j does. And: Whenever a causal model $M_k = \langle V_k,E_k,P_k\rangle$ lies on such a directed path from M_i to M_j, then M_k represents a causal structure on a level between M_i and M_j. So what we basically get by drawing a level graph is a strict order among causal models of an MLCM.

Let me now illustrate the MLCM approach for modeling mechanisms by an abstract example. Figure 6 shows the causal structures of the causal models of an MLCM plus the MLCM's level graph that connects these models and provides information about the hierarchical order of the mechanism's levels the MLCM represents. The lowest level causal model M_1's graph is $X \leftrightarrow Y \leftarrow Z \leftarrow U \rightarrow W$. One gets the higher-level model M_2 with graph $X \quad Z \leftarrow U \rightarrow W$ by marginalizing out Y, and the higher-level model M_3 with graph $Y \leftarrow Z \leftrightarrow W$ by marginalizing out X and U. Note that the MLCM's level graph does not provide any information about whether these two models (i.e., M_2 and M_3) represent structures at the same or at different levels of organization. By marginalizing out U from M_2, one arrives at the higher-level causal model M_4 with structure $X \quad Z \leftrightarrow W$. Note that the formalism again does not provide any information about whether M_4 represents a mechanism at the same level as the one represented by M_3 or not. One can further restrict M_3 and M_4 to M_5 with causal graph $Z \leftrightarrow W$. M_5 describes the represented mechanism at the top level. Note that the MLCM's level graph tells us that causal models M_2, M_3, and M_4 describe the mechanism's causal structure on levels

between the mechanism's top and its lowest level represented by M_5 and M_1, respectively.

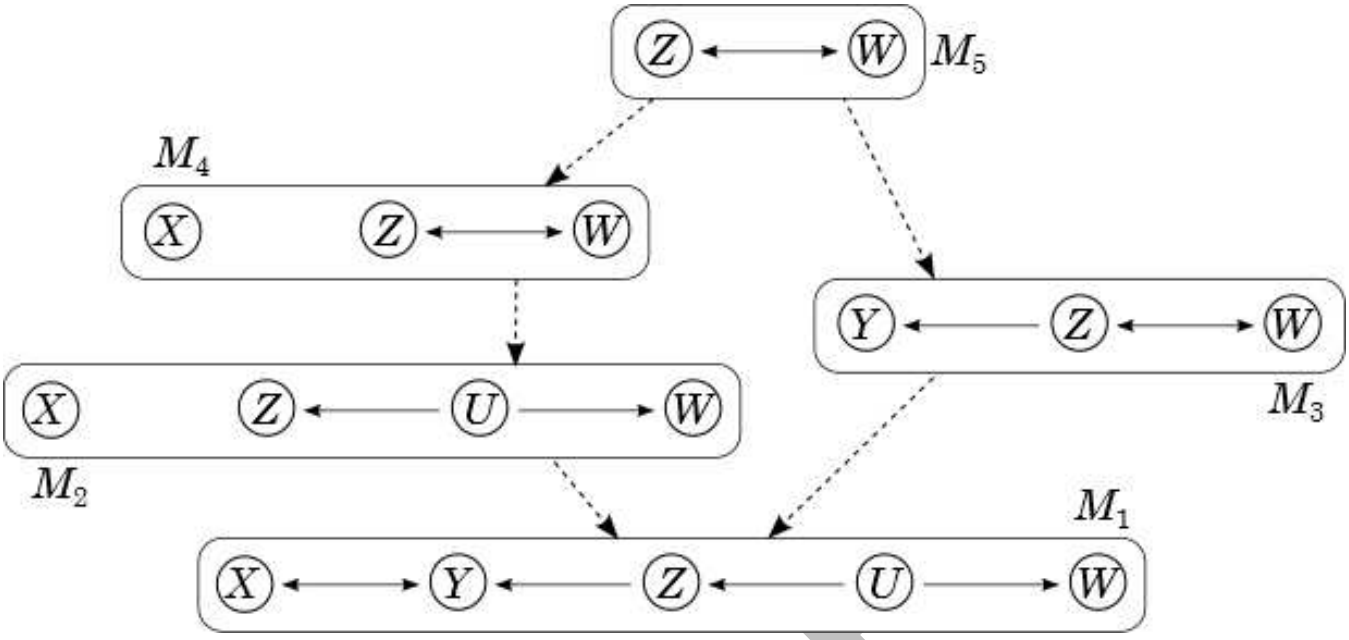

Figure 6: The graphs in the boxes are the associated causal graphs of an MLCM's causal models $M_1,\ldots,M_5$. The dashed lines are the edges of this MLCM's level graph.

As a last step I will demonstrate on our exemplary mechanism introduced in sec. 2, *viz.* the water dispenser, that MLCMs do not share problems (i) and (ii) Casini *et al.*'s (2011) RBN approach has to face and that MLCMs nicely capture another important feature of nested mechanisms:

(9) As long as the details of a mechanism are not considered, the same input should lead to the same output on all of the mechanism's levels.

The water dispenser mechanism can be represented by an MLCM $\langle \square_{\square}\square\square\langle V_1,E_1,P_1\rangle,\square_{\square}\square\square\langle V_2,E_2,P_2\rangle\rangle$, where the graph in the upper box in figure 7 shows M_2's and the one in the lower box shows M_1's causal structure. M_1 represents the water dispenser's submechanism, *viz.* the water temperature regulation unit. Note that this submechanism is not represented by a variable like in Casini *et al.*'s (2011) RBN approach, but by M_2's graph $T\rightarrow W\leftarrow B$ at the higher level. T and B are this submechanism's input variables, W is its output variable. The MLCM $\langle \square_{\square}\square\square\langle V_1,E_1,P_1\rangle,\square_{\square}\square\square\langle V_2,E_2,P_2\rangle\rangle$'s level graph is $\square_{\square}\rightarrow\square_{\square}$. When we go from M_2

to M_1, we zoom into the micro structure of the submechanism represented by $T \rightarrow W \leftarrow B$ at the top level. Since M_2 is a restriction of M_1 in the MLCM, it follows from (6)(b) that $P_1(w|t,b) = P_2(w|t,b)$ holds for arbitrarily chosen *W*-, *T*-, and *B*-values *w*, *t*, and *b*, respectively. So as long as only the variables contained in both causal models' variable sets are considered, the same input will lead to the same output at both levels, and thus, (9) is satisfied.

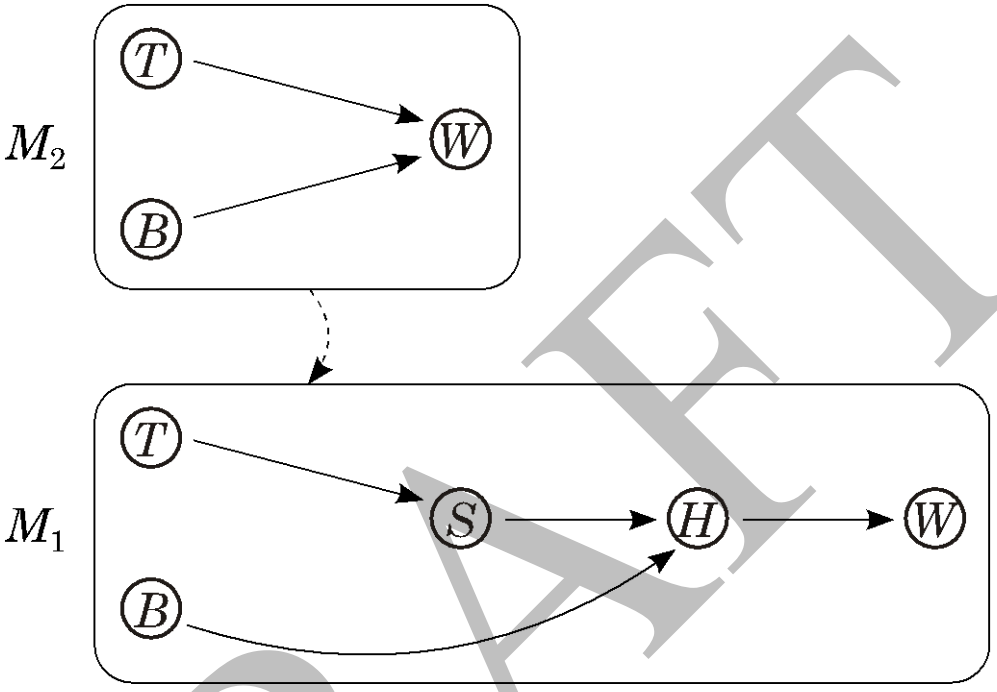

Figure 7: Representation of the water dispenser mechanism by means of a two stage MLCM. The graph with the dashed edge connecting the MLCM's two causal models M_1 and M_2 is the MLCM's level graph.

Since the causal arrows in M_1 tell us exactly how the submechanism's components *S* and *H* are causally connected to the rest of the mechanism (i.e., *T*, *B*, and *W*), the MLCM representation captures property (i): The MLCM can graphically represent the causal connections between the represented mechanism's macro and micro variables. This gives us causal information that is crucial for questions concerning explanation, manipulation, and control. It tells us why certain inputs (i.e., conditionalizing on certain *T*- and *B*-values) bring about (or explain) certain outputs (i.e., probabilities of certain *W*-values): *T* is directly causally relevant for *S*. *B* and *S* are direct causes of *H*, and *H* is the only direct cause of *W* in our toy mechanism. This causal information does tell us, for example, why *T*'s probabilistic influence on *W* breaks down when $B = 0$. It is *because* the only productive causal path from *T* to *W* goes through *H*. $B = 0$ fixes *H*'s value and, thus, probability propagation between *T* and *W* along this path is blocked when *H*'s value is fixed. It also tells us

that T's influence on W can be amplified or decreased by manipulating S or H by means of soft interventions, while B's effect on W can only be modified by changing H's behavior. The MLCM can also capture property (ii): Intervening on the mechanism's micro structure, i.e., on S or H, will typically have a probabilistic influence on the mechanism's macro behavior, i.e., on certain W-values.

Like Casini *et al.*'s (2011) RBN approach, the MLCM representation provides a unique probability distribution over the set of all variables appearing in the causal models of the MLCM. Since the first causal model $M_1 = \langle V_1, E_1, P_1 \rangle$ in an MLCM's ordering $M_1,\ldots,M_n$ also contains all variables of the causal models M_i appearing later in that particular ordering, said unique probability distribution is M_1's probability distribution P_1. When it comes to quantitative prediction, one can, thanks to the fact that every causal model appearing later in the ordering $M_1,\ldots,M_n$ is a restriction of M_1, just choose one of the causal models in the MLCM that contains all the variables of interest and then compute the probabilities for the phenomena of interest accordingly.

5. Conclusion

In this paper I tackled the question of how mechanisms, and especially their hierarchic organization, can be represented within a causal graph framework. In sec. 2 I discussed an approach for modeling such nested mechanisms proposed by Casini *et al.* (2011). I introduced Bayesian networks and recursive Bayesian networks and explained how they can be used for causal modeling. I then illustrated Casini *et al.*'s RBN approach, which suggests representing submechanisms by network variables of an RBN, by means of a very simple toy example, *viz.* the water dispenser mechanism. In sec. 3 I illustrated two problems with the RBN approach by means of the exemplary mechanism introduced in sec. 2: (i) An RBN does not graphically encode information about how a mechanism's submechanisms are causally connected to the rest of this mechanism. Such information is, however, relevant when it comes to questions of mechanistic explanation, manipulation, and control. (ii) It follows from the RBN approach that intervening on some of a mechanism's micro variables cannot have any probabilistic influence on some of this mechanism's macro behavior whatsoever. This consequence stands in stark contrast to scientific practice; scientists typically carry out so-called bottom-up experiments to distinguish between a mechanism's constitutively relevant and its irrelevant

parts. In section 4 I developed an alternative modeling approach for nested mechanism: the MLCM approach. This approach represents submechanisms not by means of a causal model's variables, but by the edges of its associated graph. I finally demonstrated, again on the exemplary mechanism of the water dispenser, that the MLCM approach does not fall prey to problems (i) and (ii) Casini *et al.*'s RBN approach has to face.

Acknowledgements: This work was supported by DFG, research unit *Causation* | *Laws* | *Dispositions* | *Explanation* (FOR 1063). My thanks go to Lorenzo Casini, Stuart Glennan, Jens Harbecke, Phyllis Illari, Marie I. Kaiser, Gerhard Schurz, Paul Thorn, Matthias Unterhuber, Ioannis Votsis, and Jon Williamson for their input and important discussions. Thanks also to Christian J. Feldbacher, Sebastian Maaß, Alexander G. Mirnig, and Lucia M. Pichler as well as to two anonymous referees for constructive criticism on an earlier version of the paper.

References

Bechtel, W., & Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Studies in the History and Philosophy of the Biological and Biomedical Sciences*, *36*, 421-441.

Bechtel, W. (2007). Reducing psychology while maintaining its autonomy via mechanistic explanations. In M. Shouten & H. L. De Joong (Eds.), *The matter of the mind: Philosophical essays on psychology, neuroscience and reduction* (pp. 172-198). Blackwell.

Casini, L., McKay Illari, P., Russo, F., & Williamson, J. (2011). Models for prediction, explanation and control: recursive Bayesian networks. *Theoria, 70*, 5-33.

Eberhardt, F., & Scheines, R. (2007). Interventions and causal inference. *Philosophy of Science*, *74*, 981-995.

Glennan, S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, *44*, 49-71.

Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for the Philosophy of Science*, *2*, 119-135.

Craver, C. (2007a). Constitutive explanatory relevance. *Journal for Philosophical Research*, *32*.

Craver, C. (2007b). *Explaining the brain*. Oxford: Clarendon Press.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67,* 1-25.

Pearl, J. (2009). *Causality*. Cambridge: Cambridge University Press.

Reichenbach, H. (1956): *The Direction of Time.* Berkeley, University of California Press.

Salmon, W. (1984). *Scientific explanation and the causal structure of the world.* Princeton: Princeton University Press.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search.* Cambridge: The MIT Press.

Steel, D. (2005). Indeterminism and the causal Markov condition. *British Journal for the Philosophy of Science, 56,* 3-26.

Williamson, J., & Gabbay, D. (2005). Recursive causality in Bayesian networks and self-fibring networks. In D. Gillies (ed.), *Laws and models in the sciences* (pp. 223-245). London: King's College Publications.

Woodward, J. (2003). *Making things happen*. Oxford: Oxford University Press.

Woodward, J. (2011). Scientific explanation. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2011 Edition). URL = http://plato.stanford.edu/archives/win2011/entries/scientific-explanation/