Another Problem with RBN Models of Mechanisms*

Alexander GEBHARTER

Received: 18/05/2015
Final Version: 2/10/2015

BIBLID 0495-4548(2016)31:2p.177-188
DOI: 10.1387/theoria.14502

ABSTRACT: Casini, Illari, Russo, and Williamson (2011) suggest to model mechanisms by means of recursive Bayesian networks (RBNs) and Clarke, Leuridan, and Williamson (2014) extend their modeling approach to mechanisms featuring causal feedback. One of the main selling points of the RBN approach should be that it provides answers to questions concerning the effects of manipulation and control across the levels of a mechanism. In this paper I demonstrate that the method to compute the effects of interventions the authors mentioned endorse leads to absurd results under the additional assumption of faithfulness, which can be expected to hold for many RBN models of mechanisms.

Keywords: recursive Bayesian networks, mechanism, modeling, intervention, manipulation, control.

RESUMEN: Casini, Illari, Russo y Williamson (2011) proponen modelar los mecanismos mediante redes bayesianas recursivas (RBNs) y Clarke, Leuridan y Williamson (2014) extienden su enfoque sobre la modelización a mecanismos que presentan retroalimentación causal. Una de las ventajas principales del enfoque RBN debería ser que proporciona respuestas a cuestiones sobre los efectos de la manipulación y el control a lo largo de los niveles de un mecanismo. En este artículo muestro que el método para computar los efectos de las intervenciones que los autores mencionados defienden conduce a resultados absurdos bajo el supuesto tradicional de fidelidad, que cabe esperar que se mantenga en muchos modelos RBN de mecanismos.

Palabras clave: Redes bayesianas recursivas, mecanismo, modelización, intervención, manipulación, control.

1. *Introduction*

In many sciences questions of explanation, prediction, and control are regularly answered by pointing at the mechanism responsible for the phenomenon of interest. Such mechanisms are typically characterized and described in a qualitative way. Glennan (1996, 52), for example, defines a mechanism underlying a behavior as "a complex system which produces that behavior by of the interaction of a number of parts according to direct causal laws". For alternative prominent characterizations of mechanisms, see, for example, Bechtel & Abrahamsen (2005, 423), and Machamer, Darden & Craver (2000, 3).

* This work was supported by Deutsche Forschungsgemeinschaft (DFG), research unit Causation | Laws | Dispositions | Explanation (FOR 1063). My thanks go to Gerhard Schurz for important discussions as well as to Alexander Christian, Christian J. Feldbacher, and two anonymous reviewers for helpful comments on an earlier version of the paper.

Casini, Illari, Russo, and Williamson (2011) argue that recursive Bayesian networks (RBNs), which were originally developed by Williamson and Gabbay (2005) to model nested causal relationships, can also be used to model mechanisms and that RBN models of mechanisms provide quantitative answers to questions concerning explanation, prediction, and control. In a follow-up paper Clarke, Leuridan, and Williamson (2014) extend the RBN approach in such a way that it can also be applied to mechanisms featuring causal feedback. One of the main selling points of the RBN approach should be that RBN models of mechanisms can be used to calculate post-intervention distributions, i.e., to predict the effects of interventions on a mechanism's parts even across levels. There are basically two possibilities to model interventions in causal models: One either represents interventions by means of so-called intervention variables. Intervention variables are variables added to the model as new causes of the variables one wants to intervene on that satisfy certain additional constraints. (For details see, e.g., Spirtes, Glymour, & Scheines 2000, sec. 3.7.2; Woodward 2003, sec. 3.1.3). Or one sets the variable one wants to intervene on to a certain value and deletes the causal arrows pointing at that variable (cf. Pearl 2009, sec. 1.3.1). Both representations can be used to compute post-intervention distributions. In (Gebharter 2014) I have argued that it is not possible to adequately represent interventions on a mechanism's micro parts as intervention variables taking certain values in RBN models of mechanisms. In particular, it follows from the RBN framework that the macro variables of an RBN are always independent of intervention variables on micro variables. According to Casini *et al.* (2011), however, interventions on micro variables which lead to a difference at the macro level are representable as a special kind of arrow breaking interventions.[1] Casini *et al.* (2011, 12f) describe how post-intervention distributions in RBN models of mechanisms can be computed in that case. In this paper I show that representing interventions on a mechanism's micro parts in RBN models of mechanisms as arrow breaking, as Casini *et al.* suggest, leads to absurd consequences under the additional assumption of faithfulness (for details see sec. 4), which can be expected to hold for many RBN models of mechanisms.

The paper is structured as follows: In section 2, I introduce Bayesian networks, causal Bayesian networks, and the notation used throughout the paper. In section 3, I briefly present Casini *et al.*'s (2011) RBN approach for modeling mechanisms. I illustrate their approach by means of an abstract example, which I will also use later on in the paper. In section 4, I present the mentioned problems Casini *et al.*'s method for computing post-intervention distributions has to face. I also discuss a possible solution and show that this solution leads straight into new and not less severe problems. I conclude in section 5.

2. *Bayesian networks and causal Bayesian networks*

A Bayesian network (BN) is a triple $\langle \mathbf{V},\mathbf{E},P\rangle$, where $\mathbf{G} = \langle \mathbf{V},\mathbf{E}\rangle$ is a directed acyclic graph (DAG) and P is a probability distribution over $\mathbf{V}$. A graph $\mathbf{G} = \langle \mathbf{V},\mathbf{E}\rangle$ consists of a set

[1] Note that in ordinary causal models it does not matter which representation of interventions one chooses, since they both lead to the same consequences. Casini *et al.*'s (2011) arrow breaking interventions, however, significantly differ from the ordinary and well-known standard arrow breaking interventions introduced by Pearl (2009). For details, see sec. 4.

V of so-called vertices (which are random variables[2] in case of a BN) and a binary relation **E** on **V**. In case of a directed graph, **E** is asymmetric. The elements of **E** are called the graph's edges and are graphically represented by arrows. Hence, "X→Y" stands short for "$\langle X,Y\rangle \in$ **E**". A chain of arrows X—...—Y connecting X and Y is called a path between X and Y.[3] A graph is acyclic if it does not feature a path of the form X→...→X.

In case of an edge X→Y, we call X the arrow's tail and Y its head. The set of variables X with X→Y in a graph is called the set of Y's parents **Par**(Y). A path of the form X→...→Y is called a directed path from X to Y. The set of all Y with X→...→Y is called the set of X's descendants **Des**(X). For a triple $\langle$**V**,**E**,$P\rangle$ to be a BN it is required that it satisfies the Markov condition (MC) (cf. Spirtes *et al.* 2000, p. 11):

Definition 1 (Markov condition)

$\langle$**V**,**E**,$P\rangle$ satisfies the Markov condition if and only if $Indep(X,\mathbf{V}\backslash\mathbf{Des}(X)|\mathbf{Par}(X))$ holds for all $X \in$ **V**.[4]

If $\langle$**V**,**E**,$P\rangle$ satisfies MC, then its graph determines the following Markov factorization for P over $\mathbf{V} = \{X_1,...,X_n\}$ (cf. Spirtes *et al.* 2000, 12):[5]

$$P(x_1, ..., x_n) = \Pi_i\, P(x_i|\mathbf{par}(X_i)) \tag{1}$$

The arrows and paths of a BN $\langle$**V**,**E**,$P\rangle$ can be causally interpreted. In that case, X is called a direct cause of Y (and Y a direct effect of X) w.r.t. variable set **V** if X→Y in **G** = $\langle$**V**,**E**$\rangle$. X is called a (direct or indirect) cause of Y (and Y an effect of X) if X→...→Y in **G** = $\langle$**V**,**E**$\rangle$. A variable $Z \neq X,Y$ lying on a directed path X→...→Y is called an intermediate cause on this path. A variable Z on a path X←...←Z→...→Y is called a common cause of X and Y, provided no variable appears more often than once on the path X←...←Z→...→Y. Finally, a variable Z on a path X—...→Z←...—Y is called a collider on this path. If the form of such a collider path between X and Y is X→...→Z←...←Y, then Z is also a common effect of X and Y.

For causally interpreted BNs, MC becomes the causal Markov condition (CMC) (cf. Spirtes *et al.* 2000, 29):

Definition 2 (causal Markov condition)

$\langle$**V**,**E**,$P\rangle$ satisfies the causal Markov condition if and only if every $X \in$ **V** is probabilistically independent of all its non-effects conditional on its direct causes.

MC is equivalent with Pearl's (2009, sec. 1.2.3) *d*-separation criterion (Lauritzen, Dawid, Larsen, & Leimer 1990). For this paper I use the following *d*-connection condi-

[2] Capital letters "X_1, ..., X_n" etc. stand for random variables, while "x_1, ..., x_n" stand for their respective values. "$X_1 = x_1$" means that X_1 has taken value x_1. Sometimes "x_1" stands short for "$X_1 = x_1$".

[3] "X—Y" is a meta symbol for "X→Y or X←Y".

[4] $Indep(X,Y|Z)$ stands for conditional probabilistic independence and is defined as $P(x|y,z) = P(x|z)$ for all X-, Y-, and Z-values x, y, and z, respectively, given $P(y,z) > 0$.

[5] "**par**(X)" stands for an instantiation of **Par**(X).

tion (cf. Schurz & Gebharter 2016, sec. 2.3), which is also equivalent with the Markov condition:

DEFINITION 3 (*d*-connection condition)

$\langle \mathbf{V},\mathbf{E},P\rangle$ satisfies the *d*-connection condition if and only if the following holds for all $X,Y \in \mathbf{V}$ and for all $\mathbf{Z} \subseteq \mathbf{V}\backslash\{X,Y\}$: If X and Y are probabilistically dependent conditional on $\mathbf{Z}$, then X and Y are *d*-connected given $\mathbf{Z}$.

DEFINITION 4 (*d*-connection/*d*-separation)

X and Y are *d*-connected given $\mathbf{Z}$ if and only if there is a path π connecting X and Y such that no U with $\rightarrow U \rightarrow$ or $\leftarrow U \rightarrow$ as part of π is in $\mathbf{Z}$, while every collider on π is in $\mathbf{Z}$ or has a descendant in $\mathbf{Z}$.

X and Y are *d* separated by $\mathbf{Z}$ if and only if they are not *d* connected given $\mathbf{Z}$.

The equivalence between MC and the *d*-connection condition reveals the full content of CMC: If a causal model satisfies CMC, then every (conditional) probabilistic dependence is produced or can be explained by some causal connection (or *d*-connection) in the model. And vice versa: If there is no causal connection between two variables (i.e., the variables are *d*-separated by a set $\mathbf{Z}$), then they are independent conditional on $\mathbf{Z}$.

3. *The RBN approach for modeling mechanisms*

Let me now briefly present Casini *et al.*'s (2011) RBN approach for modeling mechanisms, which Clarke *et al.* (2014) also endorse. I start with the notion of a recursive Bayesian network (RBN). An RBN is a BN $\langle \mathbf{V},\mathbf{E},P\rangle$ with some $X \in \mathbf{V}$ such that the values of these variables X are BNs themselves. A variable whose values are BNs is called a network variable. A variable whose values are not BNs is called a simple variable. The set of a variable Y's direct superiors $\mathbf{DSup}(Y)$ is defined as the set of network variables X in the RBN whose values' (which are BNs) variable sets contain Y as an element. Y's superiors $\mathbf{Sup}(Y)$ are defined as the transitive closure of the direct superiority relation $\mathbf{DSup}(Y)$. The set of a variable X's direct inferiors $\mathbf{DInf}(X)$ is the set of all variables Y which are vertices of the BNs which are X's values. X's inferiors $\mathbf{Inf}(X)$ are, again, defined as the transitive closure of the direct inferiority relation $\mathbf{DInf}(X)$.

Let me illustrate the introduced RBN notation by means of the following abstract example: Assume we are interested in a mechanism and how this mechanism is connected to a certain input and output. We represent the input as a variable I and the output as O. The possible states of the mechanism are modeled as the possible values of a network variable N. Then the RBN's (causal) structure will be $I \rightarrow N \rightarrow O$. Each value n of N is a BN describing one of the possible states of the mechanism. Let us further assume that the causal micro structure of the mechanism modeled is $A \rightarrow B \rightarrow C \leftarrow D$ for all of its possible states and that only the probability distribution over $\{A,B,C,D\}$ changes from state to state.[6]

[6] This is an assumption that will hold for many mechanisms. Examples may be all kinds of artificial devices such as radios, computers, and TVs. The assumption makes the presentation of the RBN approach and

Now the mechanism could be represented by the RBN depicted in fig. 1:

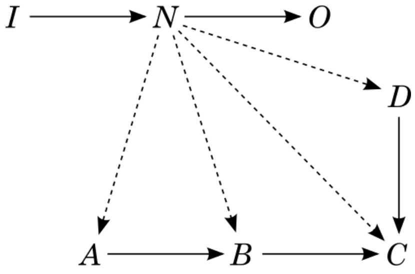

Fig. 1

The BN with graph *I*→*N*→*O* describes the system's top (or macro) level. The BNs with graphs *A*→*B*→*C*←*D* (which are the possible values of *N*) describe possible states of the mechanism represented by network variable *N*. Dashed arrows indicate relationships of direct superiority and inferiority. *N*, for example, is a direct superior of *A*, *B*, *C*, and *D*, while *A*, *B*, *C*, and *D* are direct inferiors of *N*. Casini *et al.* (2011, sec. 4) suggest to interpret the continuous arrows as representing intra-level causal relations and to interpret the dashed arrows as representing inter-level constitutive relevance relations in the sense of Craver (2007a; 2007b).

Now RBNs should allow for inter-level explanation, prediction, and for answering questions concerning the effects of manipulation and control across the levels of a mechanism. What one ultimately wants for this purpose is a probability distribution $\mathcal{P}$ over the set of all the variables $\mathcal{V} = \{I,N,O,A,B,C,D\}$. So we need to compute such a probability distribution $\mathcal{P}$ on the basis of what we have, i.e., on the basis of the BN with structure *I*→*N*→*O* and the BNs (which are the possible values of *N*) with structure *A*→*B*→*C*←*D*. For this purpose, Casini *et al.* (2011, p. 11) assume the following modeling assumption, which they call the recursive causal Markov condition (RCMC):

Definition 5 (recursive causal Markov condition)

$\langle \mathbf{V},\mathbf{E},P\rangle$ satisfies the recursive causal Markov condition if and only if $Indep(X,\mathbf{NID}(X)|\mathbf{DSup}(X) \cup \mathbf{Par}(X))$ holds for all $X \in \mathcal{V}$.

$\mathbf{NID}(X)$ is defined as the set of all $Y \in \mathcal{V}$ that are neither inferiors nor descendants of *X*. $\mathcal{V} = \{X_1, ..., X_m\}$ is the set **V** of the RBN under the transitive closure of the inferiority relation. Let $\mathbf{N} = \{\mathrm{X}_{jl},...,X_{jk}\}$ be the set of all network variables in $\mathcal{V}$. For every instantiation

the presentation of its problems much simpler and easier accessible, but nothing I do in the paper hinges on it. Note that all of Casini *et al.*'s (2011) examples feature causal micro structures with slightly different causal structures. However, their causal micro structures always represent either active or inactive states of the mechanisms represented by the corresponding network variables, meaning that they only differ in so far as the BNs representing active states feature arrows *X*→*Y* which are missing in the BNs representing inactive states. Adding these arrows to the inactive BNs does no harm. Since *X* and *Y* are connected by an arrow in some of the BNs, they will also be connected by an arrow in the BN over the set $\mathcal{V}$ of all variables at any level which is used for computing post-intervention distributions. (For details, see below.)

$\mathbf{n} = x_{jl},\ldots,x_{jk}$ of $\mathbf{N}$ one can construct a new BN: the flattening of the RBN w.r.t. $\mathbf{n}$ ("$\mathbf{n}\downarrow$" for short). The variables of such a flattening $\mathbf{n}\downarrow$ are the simple variables in $\mathcal{V}$ together with the instantiations of the network variables $\mathrm{X}_{jl}, \ldots, X_{jk}$ in $\mathbf{N}$ to their values $x_{jl},\ldots,x_{jk}$ in $\mathbf{n}$. There is a directed edge $X \rightarrow Y$ in the flattening's structure if there is an edge in the RBN's top level graph or in one of the BNs which are values of one of the network variables in $\mathbf{N}$. There is a directed edge X-->Y in $\mathbf{n}\downarrow$ if X is a direct superior of Y in the RBN. (There is no difference in probabilistic behavior of continuous and dashed edges; the only difference is the interpretation mentioned above: Continuous arrows represent intra-level causal relations and dashed arrows represent inter-level relationships of constitutive relevance.) The following figure shows a flattening of the RBN whose graph is depicted in fig. 1 w.r.t. N's value n_1:

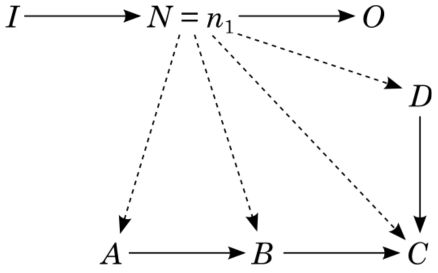

Fig. 2

The probability distribution of a flattening $\mathbf{n}\downarrow$ can be computed as follows (where X_{jl} are the direct superiors of X_i):

$$P(x_i|\mathbf{par}(X_i),\mathbf{dsup}(X_i)) = P_{xjl}(\mathbf{par}(X_i)) \tag{2}$$

Now we can compute the much sought after distribution $\mathcal{P}$ over $\mathcal{V} = \{X_1,\ldots,X_m\}$ on the basis of the flattenings $\mathbf{n}\downarrow$ as follows (where the probabilities $P(x_i|\mathbf{par}(X_i),\mathbf{dsup}(X_i))$ on the right hand side of the "=" are determined by the flattening induced by $x_1,\ldots,x_m$):

$$\mathbf{P}(x_1,\ldots,x_m) = \Pi_{\mathrm{i}}\, P(x_i|\mathbf{par}(X_i),\mathbf{dsup}(X_i)) \tag{3}$$

Since $\mathcal{P}$ over $\mathcal{V}$ factors as described in equation (3), we can construct a new BN $\langle\mathcal{V},\mathbf{E},\mathcal{P}\rangle$ with the graph $G = \langle\mathcal{V},\mathbf{E}\rangle$ depicted in fig. 1. Again, the continuous arrows can be interpreted as representing intra-level direct causal relationships, while the dashed arrows can be interpreted as inter-level relationships of constitutive relevance in the sense of Craver (2007a; 2007b). BN $\langle\mathcal{V},\mathbf{E},\mathcal{P}\rangle$ should now provide answers to questions concerning explanation and prediction as well as to questions concerning manipulation and control across the levels of the represented mechanism.

4. Troubles with interventions and RBNs

In causal models an intervention on a variable X is often represented as an arrow breaking intervention. Following Pearl (2009, sec. 1.3.1) we write "$do(x)$" for setting X's value to x

by means of an intervention. To compute post-intervention distributions $P(y|do(x))$, one deletes all the arrows in the model pointing at X and uses probabilistic independence information provided by CMC (or the *d*-separation criterion) and the resulting graph. (For details, see Pearl 1995; 2009, sec. 1.3.1.) Now Casini *et al.* (2011) adapt this method for their RBN models of mechanisms. I reconstruct their method of how to compute post-intervention distributions $\mathcal{P}(y|do(x))$ on the basis of what they do in (Casini *et al.* 2011) as follows (cf. Casini *et al.* 2011, pp. 12f). Step 1: Delete all the arrows in the BN over $\mathcal{V}$ whose head is X and whose tail is not a network variable. Step 2: Identify $\mathcal{P}(y|do(x))$ with the conditional probability $\mathcal{P}(y|x)$ computed by means of the independencies implied by RCMC and the structure of the graph resulting from step 1. Note that Casini *et al.*'s interventions —contrary to Pearl's—do not break all arrows pointing at the intervened on variables. They only break same-level causal arrows, while leaving inter-level constitutive relevance arrows intact. The main motivation for not breaking dashed inter-level arrows is that intervening on a micro variable should break the influence of its causes (at the same level), while it should still have an influence on the corresponding network variable (its direct superior) as well as on effects of this network variable at the macro level. In other words: Manipulating the lower level of a mechanism should, since the lower level constitutes the higher level, at least sometimes lead to a difference in the mechanism's macro behavior.

Let me briefly illustrate this by means of the exemplary RBN introduced in sec. 2. Assume we want to compute $\mathcal{P}(o|do(b))$ for certain O- and B- values o and b, respectively. According to step 1, we delete $A \rightarrow B$ and arrive at the structure depicted in fig. 3:

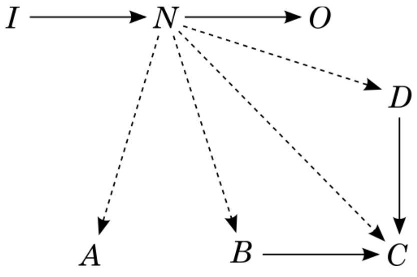

Fig. 3

According to step 2, $\mathcal{P}(o|do(b))$ should be identified with $\mathcal{P}(o|b)$ in the truncated graph, and since $\mathcal{P}(o|b) = \mathcal{P}(o,b)/\mathcal{P}(b)$, we can compute $\mathcal{P}(o|do(b))$ by computing $\mathcal{P}(o,b)$ and $\mathcal{P}(b)$ with help of the independencies implied by RCMC and the graph depicted in fig. 3 as follows:

$$\mathcal{P}(o,b) = \Sigma_i\, \mathcal{P}(n_i,\mathrm{o},\mathrm{b}) = \Sigma_i\, \mathcal{P}(o|n_i) \times \mathcal{P}(b|n_i) \times \mathcal{P}(n_i) \tag{4}$$

$$\mathcal{P}(b) = \Sigma_i\, \mathcal{P}(n_i,\mathrm{b}) = \Sigma_i\, \mathcal{P}(b|n_i) \times \mathcal{P}(n_i) \tag{5}$$

So far, so good. Now note that as long as the parameters of a BN are not fine tuned in a certain way, there will be a (conditional) dependence for every *d*-connection (cf. Schurz & Gebharter 2016, sec. 3.2). This means that most BNs will satisfy the converse of the *d*-

connection condition, i.e., they will be faithful (cf. Zhang & Spirtes 2008, p. 247; Spirtes *et al.* 2000, sec. 2.3.3). Let us assume that there are at least some mechanisms whose BNs over $\mathcal{V}$ are faithful. In that case *A* and *B* are *d*-connected over path *A<--N-->B* in the graph in fig. 3, and thus, *A* and *B* will be dependent when forcing *B* to take some value *b* by means of an intervention $do(b)$. But this means that the RBN approach implies that intervening on the effect *B* sometimes leads to a change in *A*'s probability, though *A* is a cause and not an effect of *B*. Note that because there is also a path *d*-connecting *D* and *B*, viz. *D<--N-->B*, there will also be a *B*-value *b* such that $do(b)$ does have an influence on the causally independent variable *D*. In addition, there is also a path *d*-connecting *I* and *B*, viz. *I⟶N-->B*. Thus, intervening on *B* will at least sometimes have an influence on the mechanism's macro input *I*. All of these consequences are certainly not intended by Casini *et al.* (2011).

We can generalize these findings as follows: The RBN approach leads to the absurd consequence (for faithful mechanisms) that some interventions on the effect of a micro variable will have an influence on that micro variable's non-effects (including its causes as well as the macro inputs of the mechanism represented by the corresponding network variable). This is definitely a consequence supporters of the RBN approach would like to avoid. It contradicts everything we believe to know about causation and totally blurs the distinction between observation and manipulation (cf. Pearl 2009, p. 23). It also contradicts scientific practice, since it would render any attempt to distinguish causes from effects by means of randomized experiments hopeless. We have to conclude that Casini *et al.*'s (2011) method for computing post-intervention distributions (at least in its present form) cannot give them what they want.

Here is what a supporter of the RBN approach may answer to the problems described above: In step 1 of the method for computing post-intervention probabilities $\mathcal{P}(y|do(x))$ in an RBN model of a mechanism we do not only have to delete all the arrows whose heads are *X* and whose tails are non-network variables. We also have to delete every dashed arrow whose tail is a direct superior of *X* and whose head is a non-effect of *X*; and, in addition, we also have to delete all continuous arrows whose heads are superiors of *X*.[7] According to this modified computation method, we would have to delete *A⟶B*, *I⟶N*, *N-->A*, and *N-->D* from our original BN in fig. 1. We would, thus, arrive at the structure depicted in fig. 4:

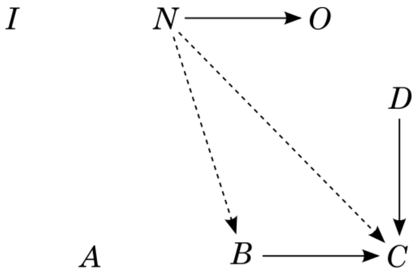

Fig. 4

[7] Note that supporters of the RBN approach cannot just simply suggest that all (i.e., continuous as well as dashed) arrows into *B* should be broken when intervening on *B*. The reason is that dashed arrows are required to allow for inter-level prediction. Without them, probabilistic influence induced by intervening on *B* could not be propagated to the variables describing the mechanism's top level.

Now the problems described above disappear. Since A and B are d-separated, there is no B-value b such that A depends on $do(b)$. Since D and B are d-separated, there is no B-value b such that D depends on $do(b)$. And, finally, since I and B are d-separated, there is no B-value b such that I depends on $do(b)$.

However, there arises a new problem not less severe. Recall that the network variable N's values should describe the possible states of the represented mechanism. When forcing B to take a certain value b by means of an intervention, we do not know in which state the mechanism is. So the causal influences $do(b)$ would have on C in the BNs which describe the mechanism's states (i.e., N's possible values) have to be weighted by the probabilities of these possible states. Because of this, the d-connecting path B<--N-->C is required to be preserved in fig. 4. But now note that one would expect that observing the values of additional micro variables will give us additional information about the probabilities of the possible states of the mechanism represented by the network variable N. If in state n_1, for example, a certain A-value a is highly probable, but a is quite improbable in all other states n_i, then learning that $A = a$ should increase the probability of $N = n_1$. In other words, $\mathcal{P}(n_1|do(b),a) > \mathcal{P}(n_1|do(b))$ should hold. But according to the modified method for computing post-intervention distributions introduced above, $\mathcal{P}(n_1|do(b),a)$ will equal $\mathcal{P}(n_1|do(b))$ simply because $\mathcal{P}(n_1|do(b),a) = \mathcal{P}(n_1|b,a)$, $\mathcal{P}(n_1|do(b)) = \mathcal{P}(n_1|b)$, and A and N are d-separated by B in the graph in fig. 4, which implies $Indep(N,A|B)$.

A similar problem arises, again, w.r.t. the mechanism's input variable I. We would expect that learning that the input is $I = i$ gives us additional information about the probabilities of the possible states of the mechanism represented by N. Thus, we would expect $\mathcal{P}(n|do(b),i) \neq \mathcal{P}(n|do(b))$ to hold for some N-, B-, and I-values n, b, and i, respectively, but the formalism implies $\mathcal{P}(n|do(b),i) = \mathcal{P}(n|do(b))$.

Summarizing, there seems to be no way out of the problem presented. When one does not delete the arrows $I \longrightarrow N$, N-->A, and N-->D from our original BN, then some interventions on B will influence some non-effects of B as well as the mechanism's input I, while deleting $I \longrightarrow N$, N-->A, and N-->D will lead to the bizarre consequence that observing the values of additional micro variables or the value of the input variable I will not give us additional information about the probabilities of the mechanism's possible states.

These findings can be generalized: It seems to be the case that the RBN approach cannot be used for computing post-intervention distributions in case of faithful mechanisms. But also its fruitfulness for getting predictions about what would happen under interventions in case of non-faithful mechanisms is more than questionable. First of all, violations of faithfulness arise only in case of parameter fine-tunings, and at least fine-tunings that produce unfaithful independencies due to cancelling paths, additional cancelling causes, or intransitive d-connections are highly improbable (cf. Spirtes *et al.* 2000, 41f; Steel 2006, 313). Thus, it can be expected that many mechanisms will be faithful. But even in case we are confronted with a non-faithful mechanism, we will get the right predictions only under very specific additional assumptions. The model's parameters must be fine-tuned in such a way that every arrow of one of the paths $I \longrightarrow N$-->B, A<--N-->B, and B<--N-->D produces some dependence between its head and its tail, but also in such a way that the three paths do not produce dependence between the variables at their end points. The latter is required to avoid the first problem discussed, i.e., that some interventions on B lead to a probability change of certain I-, A-, or D-values. The former is required to avoid the second problem presented above. It is required to guarantee that intervening on B does not screen N off from its input I as well as from A and D, and thus, that observing the values of additional micro variables can provide additional information about the probabilities of the mechanism's possible states. Summarizing, to avoid the problems already discussed by giving up faithfulness requires a very specific parameter fine-tuning. Hence, non-faithful mechanisms which can avoid the problems discussed can be expected to be even rarer than just non-faithful mechanisms.

Finally, a supporter of the RBN approach for modeling mechanisms may respond to my critique by claiming that the formalism was never intended to compute all kinds of post-intervention distributions. (Note that Casini *et al.* 2011 are not clear about which kinds of post-intervention distributions their approach should allow to compute.) She may insist that it would suffice that it is able to correctly compute the probabilities of effects of mechanisms at the macro level when manipulating the mechanism's micro parts, i.e., the effects of network variables when intervening on some of these network variables' inferiors. But even if one would be happy with a method allowing only for computing the effects of manipulations on a mechanism's parts on the mechanism's output at the macro level, there is still a problem with Casini *et al.*'s proposal. This problem arises in case there are common causes of network variables and some of their effects.

Let me illustrate this new problem, again, by means of our abstract example and assume that there is a common cause E of N and O, i.e., that the RBN's top level structure is the concatenation of $I \longrightarrow N \longrightarrow O$ and $N \longleftarrow E \longrightarrow O$. When intervening on B and not deleting the arrow N-->B (which is what Casini *et al.* 2011 originally suggested), then there are actually two active paths connecting B and O, viz. B<--$N \longrightarrow O$ and B<--$N \longleftarrow E \longrightarrow O$. Because of this we would not get the post-intervention probability $\mathcal{P}(o|do(b))$ to be expected. What we would expect when intervening on some micro parts of the mechanism represented by N is that $do(b)$ directly influences the probability distribution of N, which, in turn, only influences effects of N at the macro level. Intervening on some of the mechanism's micro parts should not lead to a probability change of a cause (such as I or E) of the network variable N. So we would expect that probabilistic influence from $do(b)$ to O is only propagated over the path B<--$N \longrightarrow O$, but not over the path B<--$N \longleftarrow E \longrightarrow O$ featuring a common cause.[8] A supporter of the RBN approach for modeling mechanisms could suggest to also delete the arrow $E \longrightarrow N$ when going from the original graph to the truncated graph used for computing post-intervention probabilities. But this move, as we already saw before, leads straightforward into new problems. In particular, learning about E's actual value would not give us any additional information about the state of the mechanism

[8] A supporter of the RBN approach could still claim that one gets the correct post-intervention probabilities when fixing the common cause E of N and O. I am indebted to an anonymous reviewer for this point. I agree that I am not able to show that post-intervention probabilities $\mathcal{P}(o|do(b),e)$ behave weird within the RBN approach. However, my argumentation still applies to post-intervention probabilities $\mathcal{P}(o|do(b))$, which in many real life cases will be more important than probabilities $\mathcal{P}(o|do(b),e)$. Assume, for example, a surgeon can only manipulate one of the micro variables, say B, directly. She cannot directly control one of the macro variables. O stands for whether a patient survives or not. In that case the surgeon and the patient are interested in the probabilities $\mathcal{P}(o|do(b))$, and not in $\mathcal{P}(o|do(b),e)$. Note that in many real life cases there will be much more common causes $E_1,...,E_n$. If only one of these E_i cannot be controlled, then there will be a path B<--$N \longleftarrow E_i \longrightarrow O$ to which my argumentation above applies.

represented by N when manipulating this mechanism's micro variable B. But, again, this would be strange. Since N's values represent the possible states the modeled mechanism can be in, conditionalizing on E should give us some hint on the mechanism's state when intervening on B.

5. *Conclusion*

Casini *et al.*'s (2011) RBN approach for modeling mechanisms, which is also endorsed by Clarke *et al.* (2014), should provide predictions about the effects of manipulations on a mechanism's parts across levels. Such manipulations are typically represented in causal models either by intervention variables or by deleting arrows and fixing the values of the manipulated variables. In (Gebharter 2014) I have shown that the former is not possible in RBN models of mechanisms. In this paper I showed that also representing interventions as a certain kind of arrow breaking interventions is highly questionable in RBN models of mechanisms. The method Casini *et al.* suggest for computing post-intervention distributions can be expected to regularly produce absurd consequences. Correct post-intervention distributions can only be computed in case the mechanism modeled is non-faithful and satisfies very specific additional conditions. Since such mechanisms can be expected to be rare, the RBN approach seems to be more or less useless for answering questions concerning the effects of manipulation and control across the levels of mechanisms, which was one of the main motivations for developing the approach in the first place.

The main result of the paper is a negative one. It leaves the question of how to model mechanisms in such a way that all kinds of post-intervention distributions can be computed unanswered. For an alternative approach that also allows to adequately compute post-intervention distributions, see (Gebharter 2014) and (Gebharter & Schurz forthcoming). For how mechanism discovery might work in such an approach, see (Murray-Watters & Glymour 2015). For recent objections to this approach, see (Casini forthcoming).

REFERENCES

Bechtel, William and Adele Abrahamsen. 2005. Explanation: A Mechanist Alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences* 36: 421-441.

Casini, Lorenzo. Forthcoming. How to model mechanistic hierarchies. *Philosophy of Science.*

—, Phyllis lllari, Federica Russo, and Jon Williamson. 2011. Models for Prediction, Explanation and Control: Recursive Bayesian Networks. *Theoria* 26: 5-33.

Clarke, Brendan, Bert Leuridan, and Jon Williamson. 2014. Modeling Mechanisms with Causal Cycles. *Synthese* 191: 1651-1681.

Craver, Carl. 2007a. Constitutive Explanatory Relevance. *Journal of Philosophical Research* 32: 3-20.

—, 2007b. *Explaining the Brain*. Oxford: Clarendon Press.

Gebharter, Alexander. 2014. A Formal Framework for Representing Mechanisms? *Philosophy of Science* 81: 138-153.

— and Gerhard Schurz. Forthcoming. A Modeling Approach for Mechanisms Featuring Causal Cycles. *Philosophy of Science.*

Glennan, Stuart. 1996. Mechanisms and the Nature of Causation. *Erkenntnis* 44: 49-71.

Lauritzen, S. L., Alexander P. Dawid, B. N. Larsen, and H. G. Leimer. 1990. Independence Properties of Directed Markov-Fields. *Networks* 20: 491-505.

Machamer, Peter, Lindley Darden, and Carl Craver. 2000. Thinking About Mechanisms. *Philosophy of Science* 67: 1-25.

Murray-Watters, Alexander and Clark Glymour. 2015. What Is Going on Inside the Arrows? Discovering the Hidden Springs in Causal Models. *Philosophy of Science* 82: 556-586.

Pearl, Judea. 1995. Causal Diagrams for Empirical Research. *Biometrika* 82: 669-688.

—. 2009. *Causality*. Cambridge: Cambridge University Press.

Schurz, Gerhard and Alexander Gebharter. 2016. Causality as a Theoretical Concept: Explanatory Warrant and Empirical Content of the Theory of Causal Nets. *Synthese* 193: 1073-1103.

Spirtes, Peter, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*. Cambridge: MIT Press.

Steel, David. 2006. Homogeneity, Selection, and the Faithfulness Condition. *Minds and Machines* 16: 303-317.

Williamson, Jon and Dov Gabbay. 2005. Recursive Causality in Bayesian Networks and Self-Fibring Networks. In *Laws and Models in the Sciences*, edited by Donald Gillies, 173-221. London: Oxford University Press.

Woodward, James. 2003. *Making Things Happen*. Oxford: Oxford University Press.

Zhang, Jiji and Peter Spirtes. 2008. Detection of Unfaithfulness and Robust Causal Inference. *Mind and Machines* 18: 239-271.

Alexander Gebharter is a research fellow at the Düsseldorf Center for Logic and Philosophy of Science (DCLPS). His research interests lie in philosophy of science and metaphysics. He is especially interested in causation and related topics such as modeling, intervention and control, mechanisms, supervenience, theoretical concepts, empirical content, etc.

Address: Department of Philosophy, Heinrich Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany. E-mail: alexander.gebharter@gmail.com