Noname manuscript No.
(will be inserted by the editor)

Modelling mechanisms with causal cycles

Brendan Clarke, Bert Leuridan & Jon Williamson

Draft of September 28, 2013

Abstract Mechanistic philosophy of science views a large part of scientific activity as engaged in modelling mechanisms. While science textbooks tend to offer qualitative models of mechanisms, there is increasing demand for models from which one can draw quantitative predictions and explanations. Casini et al. (2011) put forward the Recursive Bayesian Net (RBN) formalism as well suited to this end. The RBN formalism is an extension of the standard Bayesian net formalism, an extension that allows for modelling the hierarchical nature of mechanisms. Like the standard Bayesian net formalism, it models causal relationships using directed acyclic graphs. Given this appeal to acyclicity, causal cycles pose a *prima facie* problem for the RBN approach. This paper argues that the problem is a significant one given the ubiquity of causal cycles in mechanisms, but that the problem can be solved by combining two sorts of solution strategy in a judicious way.

1 Introduction

The concept of 'complex-system mechanism', which is commonly defined such that a mechanism's behaviour is realized by the organized behaviour of its component parts, plays an increasingly important role in philosophy of science. A natural question to ask is how mechanisms can or should be modelled. Adequately modelling mechanisms is a precondition for succesful mechanistic prediction, intervention and/or explanation. Non-formal models of mechanisms have been discussed at length in philosophy of science; see for example Glennan (2005), Bechtel and Abrahamsen (2005), and Craver (2006). Some authors have posited the need, however, to develop formal models of mechanisms that might be used to draw quantitative, as well as qualitative, inferences from the model (Lazebnik, 2002; Bechtel, 2011). In this paper, we will elaborate one possible formal approach: mechanisms can be modelled by means of Recursive

Address(es) of author(s) should be given

Bayesian Networks (RBNs). The RBN formalism is an extension of the standard Bayesian net formalism. In contrast with standard Bayesian nets, RBNs can be used to model the *hierarchical* nature of mechanisms. This approach to modelling mechanisms was originally put forward by Casini et al. (2011). One limitation of that work, however, was that it lacked a principled way of handling mechanisms that involve causal cycles. The primary aim of this paper is to provide such an account. Given the ubiquity of cycles in mechanisms (see §3), this is an important step forward in the development of the RBN approach to mechanistic modelling.

The structure of the paper is as follows. §2 and §3 together motivate our project, in that they substantiate the need for an RBN account that can handle cycles. As such, they make clear both why Casini et al. (2011) have provided an interesting approach to mechanistic modelling, *and* in what ways their account should be modified to be useful when modelling cases from scientific practice. In §2 we will highlight three important features of mechanisms as they are discussed in recent philosophy of science. It will emerge that the machinery of Recursive Bayesian Networks will be well suited to modelling mechanisms—on the condition that the acyclicity assumption, inherited from standard Bayesian networks, is dropped. In §3 we will argue that cycles are everywhere in the sciences, in particular in the biomedical and the biological sciences. We also offer a threefold classification of cycles. We then discuss three mereologically nested examples of biomedical mechanisms with cycles, drawn from recent sleep research, in some detail. In §4 we introduce the framework of Recursive Bayesian Networks and the ordinary causal Bayesian networks to which they are related. We also show how they may be used for inference (e.g., prediction). Finally, we show that the cycles discussed in §3 pose a conundrum for RBNs as defined in Casini et al. (2011), which are assumed to be acyclic. In §5 we sketch two ways to handle causal cycles in ordinary causal Bayesian nets. The first is a discussion of the extent to which well-known results (relating to *d*-separation and the Causal Markov Condition) carry over from the acyclic to the cyclic case. The second makes use of Dynamic Bayesian Nets (DBNs). Both these solutions are then incorporated in the RBN framework in §6, thus allowing for Recursive Bayesian Networks that contain cycles. Which solution to apply depends on the type of problem that is studied, as well as on pragmatic considerations, such as the granularity of analysis that is required. We distinguish between static and dynamic problems, relate them to the three-fold classification of cycles offered in §3, and provide examples of how they can be handled. Finally, in §7, we summarize our approach and make some concluding remarks.

Before we start, we would like to make two terminological and two substantive remarks. The first remark concerns the notion 'recursive'. In the older literature on structural equation modelling and causal Bayesian nets, this was often used in the sense of 'acyclic' (see, e.g., Spirtes, 1995). Hence one may worry that 'cyclic Recursive Bayesian Net' is a contradiction in terms. As RBNs use 'recursive' in a different sense—the more standard sense of a recursive or inductive definition, which can appeal to another instance of itself—this worry should not arise.

The second remark concerns the notion of 'cyclicity' itself. The topic of cyclic causality is studied in diverse domains, ranging from AI and computer science, through philosophy of science, to a wide range of empirical sciences. In these domains, several synonyms (or close proxies) for 'cyclicity' are used, among which 'bidirectionality' and 'feedback' are the most common ones. Since 'bidirectional arcs' are often used in the literature on causal discovery to denote the existence of (unobserved) confounding factors instead of bidirectional causal influences, and since 'bidirectional causality' often refers to 'atomic' cycles of the form $A \rightleftarrows B$ (where A is a direct cause of B and vice versa), we will avoid using the word 'bidirectionality'. In §3, we will touch on the notion of 'feedback' in more detail, distinguishing three types of feedback and elucidating their links with 'cyclicity'.

The third remark relates to the precise goal of our paper, which is *modelling* mechanisms. We do not intend to tackle issues of *causal* (or *mechanism*) *discovery*, such as specifying algorithms for inferring RBNs from observational and/or experimental data. We presuppose the mechanism is known (be it completely or incompletely, fallibly or infallibly) and ask how we can best model it so as to draw quantitative inferences. This account may then serve as a basis for further research on formal methods for mechanism discovery.[1]

The fourth remark concerns the interpretation of causality. In the mechanistic literature, Woodward's *interventionist* account of causality is relatively widespread (see Woodward, 2003, for the interventionist account; see e.g., Glennan, 2002, Woodward, 2002[2], Craver, 2007 and Leuridan, 2010, for appeals to the interventionist account of causality within a mechanistic framework). As is well known, Woodward's account nicely fits the causal Bayesian nets literature.[3] Another interpretation that has been proposed with an eye to causal Bayesian nets is the *epistemic* account, according to which causation is a feature of the way we represent the world rather than the world itself, yet it is objective in the sense that if two agents with the same evidence disagree regarding a causal claim, one may be right and the other wrong; see Williamson (2005, chapter 9). In this paper, we will not adopt a specific account of causality. Any account that suits the causal Bayesian nets framework, such as the two just mentioned, can be chosen.

[1] Note that, as with all models, a recursive Bayesian network model only models some aspects of a mechanism. The main goal is to model the hierarchical structure of the mechanism together with the causal structure at each level of the hierarchy, in such a way that the model can be used to draw quantitative inferences. See Casini et al. (2011) for a fuller presentation of the motivation behind this sort of model, and §7 of this paper for pointers to possible limitations of the RBN approach.

[2] Woodward's concept of 'mechanism', or more precisely: of 'mechanistic model', is not explicitly multi-level or hierarchical, in contrast to those on which we focus in this paper. In the next section, the hierarchical nature will serve as one of the main reasons to adopt the RBN approach to mechanistic modelling.

[3] As such, his account of causality also forms the starting point for causal Bayes net accounts of the structure of scientific theories (see Leuridan, 2014).

In the interest of readability, we aim to keep the paper as non-technical as possible. Further technical details concerning the frameworks we use can be found in the references.

2 Importance for philosophy of science

The concept of 'complex-system mechanism' plays an increasingly important role in philosophy of science. In this paper, we will not survey the overwhelming literature on mechanisms (see for example Machamer et al., 2000, Glennan, 1996, Glennan, 2002, Bechtel and Abrahamsen, 2005). Rather, we will focus on two recent works in the mechanistic tradition, one by Carl Craver and one by William Bechtel, and focus on three key features of mechanisms and mechanistic models they discuss. These three features will set the agenda for our paper.

In his book *Explaining the Brain*, Craver (2007) gives a very detailed account of mechanisms. A first feature that emerges from his work, is that mechanisms are *hierarchically organized.* As we wrote above, mechanisms are commonly defined such that their higher-level behaviour is realized by the organized lower-level behaviours of their component parts. This hierarchical structure need not be confined to two levels. The behaviours of a mechanism's components may themselves be mechanistically explicable as well. In fact, there may be a whole series of nested mechanisms (see Craver, 2007, 188–195). In a recent paper, "Mechanism and Biological Explanation," William Bechtel expresses a similar view: mechanistic explanations are always multilevel accounts, focusing on the mechanism's parts, operations and organisation, on the phenomenon exhibited by the whole mechanism, and on the mechanism's environment (Bechtel, 2011, 538).

This hierarchical structure of mechanisms does not require, however, that our descriptions of such hierarchies be open-ended in the downwards direction. Models typically *bottom-out* in lowest-level mechanisms which are accepted as relatively fundamental or unproblematic (in a given context); see Machamer et al. (2000, 13) and Craver (2007, 193).

The hierarchical structure of mechanisms is illustrated in figure 1 (which is adapted from Craver, 2007, 7). The X's are components in the mechanism for S's ψ-ing. The ϕ's are their respective behaviours. Solid arrows represent intra-level causal relations. The dotted lines denote inter-level constitutive relevance.

Figure 1 brings us to a second important feature of mechanisms: their possibly *cyclic* causal organization. X_2 and X_3 are cyclically causally connected: X_2 is a cause of X_3, and vice versa (strictly speaking, X_2's ϕ_2-ing is a cause of X_3's ϕ_3-ing and vice versa, but let us omit such circumlocutions). This is not a mere coincidence or a slip of the pen on Craver's behalf. Elsewhere, he writes that symmetric causal relations exist, although he sometimes seems to underrate their incidence (e.g. Craver, 2007, 153). He also mentions cases of

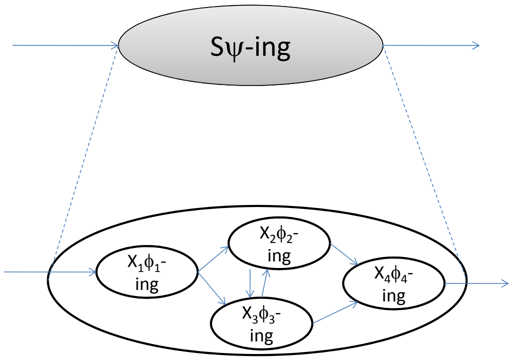


Fig. 1 A phenomenon (top) and its mechanism (bottom)

causal feedback in the neurosciences (the discipline of interest in his book) on several occasions.[4]

Bechtel attaches great importance to cyclicity. A crucial feature of biological organisms is 'their autonomy—their ability to maintain themselves as systems distinct from their environment by directing the flow of matter and energy so as to build and repair themselves' (Bechtel, 2011, 535). For this autonomy, cyclic organization is pivotal: 'Autonomous systems must employ a nonsequential or cyclic organization such as negative feedback . . . ' (Bechtel, 2011, 544). Mechanisms cannot be modelled merely sequentially without loss of crucial information.

A third important feature, which is heavily stressed by Bechtel (2011, 536–537), is that to adequately model a mechanism, one has to model not only its qualitative aspects, but also its *quantitative aspects*. Bechtel proposes computational modeling and dynamic systems analysis as methods to account for the non-qualitative aspects of mechanisms. We will propose a different approach here: an approach based on causal Bayesian nets.

As we shall see in §4, causal Bayesian nets can model both the qualitative aspects of causal structures by means of a causal graph *and* their quantita-

[4] For mentions of causal feedback in *Explaining the Brain*, see e.g. pages 81 and 178-180. Several of Craver's figures also contain cycles: see figure 4.1 (p. 115) and figure 4.6 (p. 121) and relatedly 5.7 (p. 189) and 5.8 (p. 194). Moreover, figure 3.2 (p. 71) and relatedly 4.1 (p. 166), leave open the possibility of causal feedback.

tive aspects by means of the associated probability distribution. Hence this approach would automatically meet Bechtel's call for a quantitative account of mechanisms. Following Casini et al. (2011), we use Recursive Bayesian Nets (RBNs) instead of standard causal Bayesian nets so as to account for the mechanisms' hierarchical organization (§4). In order to account for the mechanisms' possibly cyclic causal organisation, we will explore existing solutions to the problem of cyclicity in causal Bayesian nets (§5) and incorporate these in the RBN framework (§6). As a result, we provide a formal account of mechanisms that combines all three features discussed above.

But first we shall explore the abundance of cyclic mechanisms in the sciences and distinguish between several types of causal cycles, as this will influence the choice of technical solution in each context (§3).

3 Importance for the sciences

Although not every causal relationship of interest to the sciences exhibits cyclicity, very many causes of practical importance do. A bibliographic search in the ISI Web of Science for "*topic=(causal feedback)*" on the 4th of November 2011 yielded 1,161 hits in disciplines as diverse as cell biology, biochemistry, molecular biology, neuroscience, environmental studies, psychology and social psychology, while a wider search in all of the ISI's databases yielded a total of 1,603 hits. Cycles are everywhere in the sciences.

They are particularly prevalent in the biomedical and biological sciences. Examples include metabolic cycles (such as Krebs' cycle), organismal life cycles (such as the malaria-causing organisms of the genus *Plasmodium*), homeostatic pathways (such as blood glucose regulation) and pathological processes. A survey of the 790 images contained in a recent medical textbook, *Davidson's Principles & Practice of Medicine*, 20th edition (Boon et al., 2006), revealed a total of 154 images that contained some graphical representation of causal processes. 51 of these figures (33%) were at least partially cyclic, conveying knowledge about the regulation of the cell cycle, the life-cycles of various pathogenic organisms, the homeostasis of fat, fluids and ions, the arrangement of hormone systems and the development of disease.

A simple example of this kind of cycle is post-traumatic raised intracranial pressure. Here, trauma to the head may cause swelling of the brain. This swelling increases the pressure within the fixed volume of the skull. The consequence of this increased intracranial pressure is to reduce cerebral perfusion, which in turn causes cerebral hypoxia. This hypoxic insult causes damage to the brain cells, which leads to further swelling.

One interesting feature of these causal cycles concerns the way that the organisation of parts and operations governs the type of feedback seen in that cycle. Three arrangements are possible. *Negative-feedback cycles* are those in which the properties of the parts in the cycle tend to maintain the *status quo* by virtue of the organisation of their operations. Thus, the higher level phenomena produced by negative-feedback cycles tend to be stable, as in the case of many metabolic and homeostatic processes.[5] With respect to medicine, this means that negative-feedback cycles are typically physiological, rather than disease-producing. Second are *positive-feedback cycles*. Here, the organisation of parts and operations tends to produce divergence from equilibrium of one or more parts over time. These kinds of cycles are typically associated with the production of identifiable disease states, such as the head trauma example given above.[6] In this case, the degrees of swelling, intracranial pressure, and cellular damage will tend to increase, while cerebral perfusion and oxygenation will tend to decrease. The higher level phenomena produced by positive-feedback cycles thus demonstrate exponential growth over time—at least until restrained by external factors. Finally, a third kind of cycle exists in which the organisation of the cycle neither tends to produce movement towards, or away from, equilibrium. In these kinds of *contingent-feedback cycles*, the actual direction of change is predominantly governed by factors extrinsic to the cycle. For example, the parts of a parasitic life-cycle are largely governed by factors external to that cycle, meaning that either positive or negative feedback may occur in different instantiations.[7]

The type of feedback seen in a particular cycle partly depends on the manner in which we investigate that cycle. For example, an oscillating system studied over durations much shorter than its period may appear to demonstrate positive feedback, yet will appear to show negative feedback if studied at much longer durations. An example of this granularity in the description of feedback can be seen in the pathway by which the concentration of thyroid hormones is maintained (see figure 2). The secretion of thyrotropin-releasing hormone (TRH) from the hypothalamus stimulates the pituitary gland to secrete thyroid-stimulating hormone (TSH). In turn, this causes the thyroid to secrete the hormone thyroxine.[8] The resultant increase in circulating thyroxine levels inhibits the secretion of TRH by the hypothalamus, which has the effect, via reduced TSH secretion, of reducing the concentration of thyroxine. When viewed over the long-term, this is a negative-feedback cycle, as the concentrations of each of the hormones involved tend towards equilibrium. However, at very short durations, individual parts of the mechanism may undergo changes away from their equilibrium point. Thus, the type of feedback

[5] This property of negative feedback systems to tend toward equilibrium is the case when there is no significant delay in the system. When delay is present, as it is in many biological systems, oscillations will tend to arise. We would like to thank Mike Joffe for pressing us on this issue.

[6] However, this is not always the case. For instance, various positive feedback loops in pregnancy serve to appropriately maintain hormone levels.

[7] These kinds of contingent-feedback cycles are, in other words, more sensitive to background conditions than the other two kinds. This makes them *unstable*, in Mitchell's sense of stability as describing the sensitivity of relations to their background conditions (Mitchell, 2009, 56). This should be discriminated from *robustness*, which describes the degree to which a function is maintained when one or more constitutive elements are disrupted (Mitchell, 2009, 69-73).

[8] This cycle is more complicated than suggested above. For example, the thyroid secretes two hormones—T3 and T4—which can be interconverted, and feedback occurs at various intermediate points in the cycle. But this simple version is adequate for our discussion.

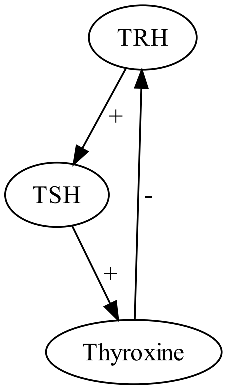


Fig. 2 Thyroxine example. TRH: thyrotropin-releasing hormone; TSH: thyroid-stimulating hormone.

modelled depends on the granularity with which we investigate a phenomenon of interest. Modelling feedback, as with many other considerations in mechanistic modelling, also depends on pragmatic factors such as the required level of detail.

Given this brief introduction to cycles in practice, the remainder of this section will discuss three mereologically nested examples of biomedical mechanisms with cycles, drawn from recent work in sleep research.

3.1 Public health example

Insufficient sleep is correlated with mortality (Grandner et al., 2010). However, the mechanism underlying this association is highly complex and poorly understood. As figure 3 suggests, an extensive network of social, psychological and pathological states causally interact with both sleep and mortality. The duration and quality of sleep interact cyclically with a range of mortality-associated physiological and social outcomes including obesity, cardiovascular disease, stress and metabolic dysfunction (indicated by edges **C** and **E**, figure 3). To illustrate this, consider the relationship between insufficient sleep and cardiovascular disease. Broadly, while cardiovascular disease causes impaired sleep (perhaps by causing chest pain or shortness of breath while lying down), impaired sleep may also cause cardiovascular disease (perhaps by increasing blood pressure).

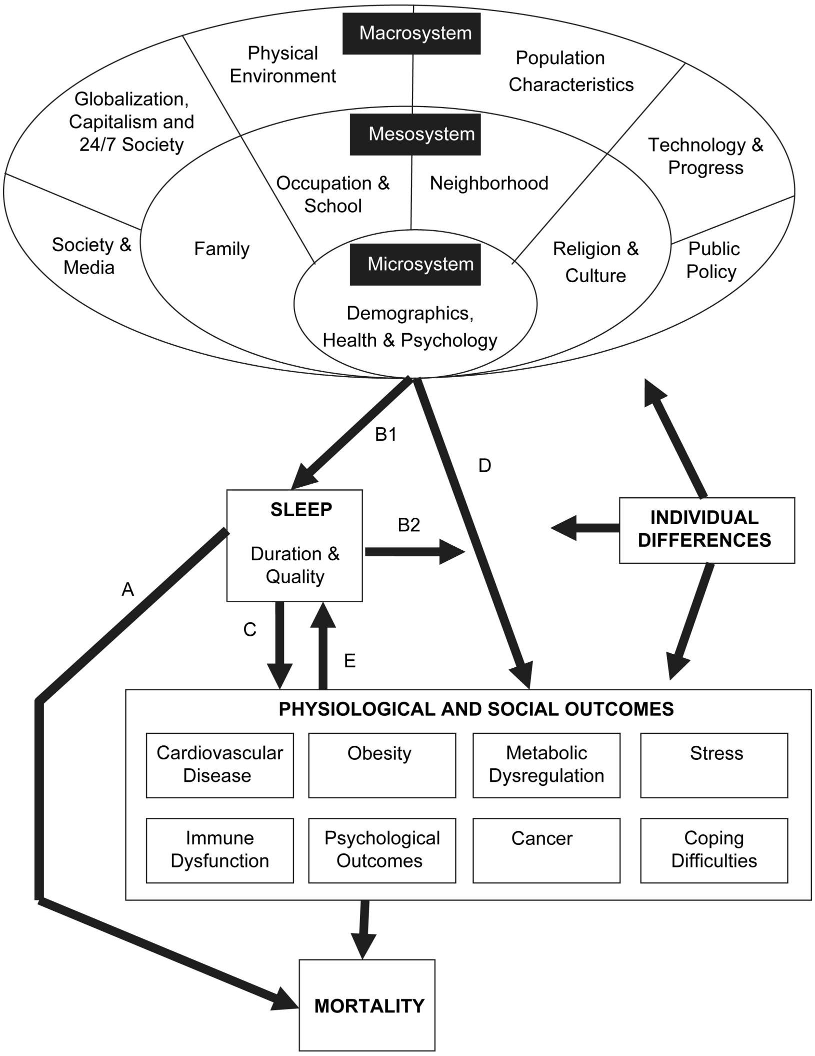


Fig. 3 Figure showing cyclical interactions between sleep and health outcomes (Grandner et al., 2010, 200). Reprinted from *Sleep Medicine Reviews*, **14**(3), Michael A. Grandner, Lauren Hale, Melisa Moore and Nirav P. Patel, Mortality associated with short sleep duration: The evidence, the possible mechanisms, and the future, pages 191-203. Copyright 2010, with permission from Elsevier.

3.2 Clinical example

One way in which sleep and cardiovascular disease interact is by the clinical syndrome known as Obstructive Sleep Apnoea (OSA). This is a condition where excessive relaxation of the tissues of the throat leads to occlusion of the upper airway, temporarily but completely interrupting breathing during sleep. This transient suffocation then leads to an arousal event, where the individual

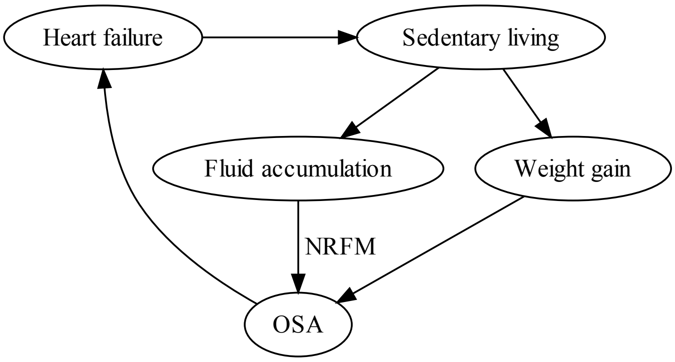


Fig. 4 Diagram showing the cyclical relationship between sleep apnoea and heart failure. NRFM: nocturnal rostral fluid movement; OSA: obstructive sleep apnoea (Gottlieb et al., 2010; Yumino et al., 2010).

experiences a complex set of cardiorespiratory responses, terminated in wakefulness and the re-establishment of normal breathing. These stressful arousal events may occur on dozens of occasions during each period of sleep.

Many diseases are correlated with OSA. One example is heart failure. Here, the heart is unable to pump sufficient blood to keep pace with normal circulatory demands. While the aetiology of heart failure is complex, one known cause of it is OSA (Gottlieb et al., 2010), which acts by a combination of short- and long-term mechanisms. For instance, in the short term, arousal events dramatically increase both blood pressure and cardiac oxygen demands. This may produce gradual cardiac remodelling, and excessive sympathetic nervous system activity, leading to an increased chance of developing heart failure (McNicholas and Bonsignore, 2007, 161).

However, as figure 4 suggests, not only is OSA a cause of heart failure, but it may also result from it. First, heart failure—in common with many chronic diseases—often causes sufferers to be very tired, leading to the adoption of a highly sedentary life-style. Fairly intuitively, being sedentary tends to cause weight gain, which can predispose to OSA by increasing the size of the neck. Second, heart failure can cause OSA via a pathological process known as nocturnal rostral fluid movement (NRFM). The basic idea is this. One of the consequences of heart failure is a condition known as dependent oedema, which is characterised by the abnormal accumulation of extracellular fluid in the lower parts of the body (the ankles, for instance). In NRFM, lying down (while, for instance, sleeping) causes this fluid to migrate up from the lower part of the body towards the head and neck. Here, the fluid accumulates in the soft tissues of the throat, producing an enlargement of the soft tissues in an analogous way to obesity, and similarly increasing the chances of airway obstruction (Yumino et al., 2010). Heart failure, via both weight gain and NRFM, causes OSA, while OSA, via a series of complex processes associated with arousal events, causes heart failure.

3.3 Neuroscience example

A neurologically important feature of normal NREM sleep is the slow wave-like rhythm in which neurones across the cortex 'beat' at a frequency of about 1Hz.[9] This slow wave sleep is thought to be causally significant in consolidating new memories (Marshall et al., 2006). This oscillation comes about from a cycle that obtains between three populations of neurones (Crunelli and Hughes, 2010), as described in figure 5. First, two populations of neocortical neurones—'a subset of pyramidal neurons in layers 2/3 and 5 and a group of Martinotti cells that is exclusively located in layer 5' (Crunelli and Hughes, 2010, 11)—and a group of thalamic cells, known as cortically projecting thalamic (CT) neurones, act as intrinsic pacemakers, spontaneously generating a slow oscillation. Together, the effect of these pacing cells is to stimulate two populations of nerves in the thalamus—the CT and nucleus reticularis thalami (NRT) neurones. In turn, these thalamic cells, once stimulated in this way, evoke a strong oscillatory response from the thalamus more broadly. This has the effect of stimulating 'virtually all cortical neurones' (Crunelli and Hughes, 2010, 10) to produce the <1Hz oscillation seen on EEG. This waveform therefore arises by virtue of the cyclical interactions between three different populations of neurones, as Crunelli and Hughes suggest: 'the full EEG manifestation of the slow rhythm requires the essential dynamic tuning provided by their complex synaptic interactions' (Crunelli and Hughes, 2010, 14).

4 Recursive Bayesian networks

In this section we will introduce Recursive Bayesian networks and see that causal cycles present a *prima facie* problem for this formalism.

4.1 Origins

Bayesian networks were developed in the 1980s in order to facilitate, among other things, quantitative reasoning about causal relationships (Pearl, 1988).[10]

[9] Similar examples of neurological oscillators are also discussed by Bechtel (2011, 548-549).

[10] Structural equation modelling had previously been put forward for this purpose. But structural equation models attempt to model deterministic relationships between cause and effect (with error terms which are usually assumed to be independent and normally distributed), while Bayesian networks seek to represent the probability distribution of the variables in question. In general it is harder to devise an accurate model of deterministic relationships than it is to determine probabilistic relationships between cause and effect.

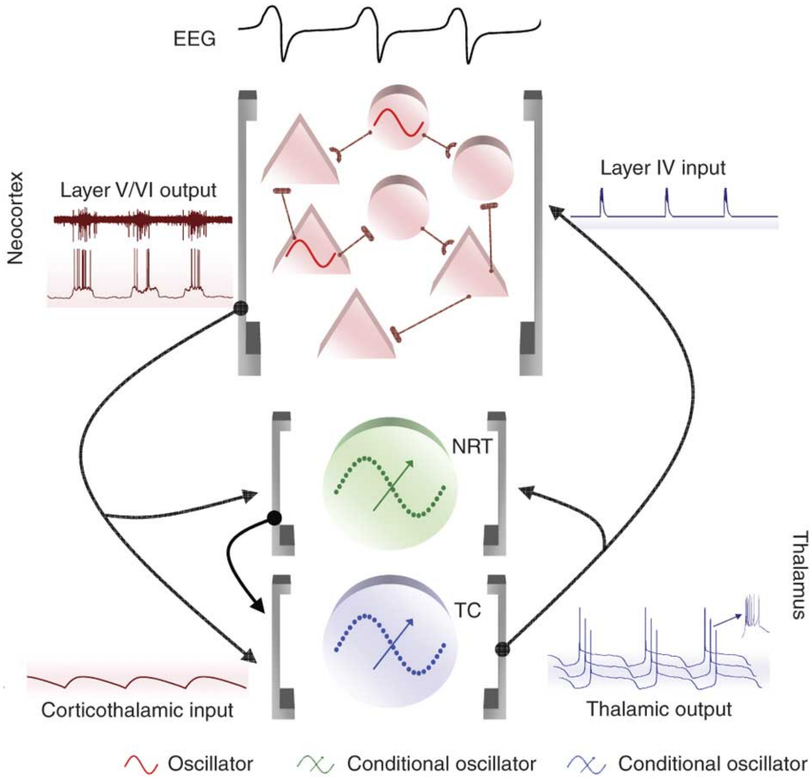


Fig. 5 Figure showing cyclical interactions between cortical and thalamic oscillators in the production of the slow (<1 Hz) rhythm. (Crunelli and Hughes, 2010, 14). Reprinted by permission from Macmillan Publishers Ltd: *Nature Neuroscience*, **13**(1), Vincenzo Crunelli and Stuart W. Hughes, copyright 2010.

A causally-interpreted Bayesian net uses a directed acyclic graph (DAG) to represent qualitative causal relationships and the probability distribution of each variable conditional on its parents to represent quantitative relationships amongst the variables. The Recursive Bayesian Network (RBN) formalism was developed to model nested causal relationships such as *[smoking causing cancer] causes tobacco advertising restrictions which prevent smoking which is a cause of cancer* (Williamson, 2005, Chapter 10). Standard Bayesian nets cannot be applied to this task because they cannot model causal relationships acting as causes, and they do not allow variables such as *smoking* to occur at more than one place in the network. Casini et al. (2011) then went on to apply the RBN formalism to the modelling of mechanisms, which are often thought of as composed of nested levels of causal relationships.[11]

[11] Note that RBNs are not the only hierarchical extension of Bayesian nets. See Williamson (2005, §10.2) for a comparison between RBNs and other related formalisms.

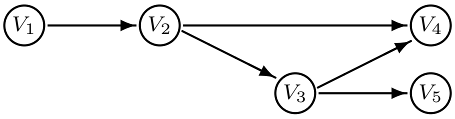


Fig. 6 A directed acyclic graph

4.2 Bayesian nets

A Bayesian net consists of a finite set $V = \{V_1, \ldots, V_n\}$ of variables, each of which takes finitely many possible values, together with a directed acyclic graph (DAG) whose nodes are the variables in V, and the probability distribution $P(V_i|Par_i)$ of each variable V_i conditional on its parents Par_i in the DAG.[12] Fig. 6 gives an example of a directed acyclic graph; to form a Bayesian net the probability distributions $P(V_1), P(V_2|V_1), P(V_3|V_2), P(V_4|V_2V_3)$ and $P(V_5|V_3)$ need to be provided. The graph and the probability function are linked by the *Markov Condition* which says that each variable is probabilistically independent of its non-descendants, conditional on its parents, written $V_i \perp\!\!\!\perp ND_i \mid Par_i$. Fig. 6 implies for instance that V_4 is independent of V_1 and V_5 conditional on V_2 and V_3. A Bayesian net determines a joint probability distribution over its nodes via $P(v_1 \cdots v_n) = \prod_{i=1}^{n} P(v_i|par_i)$ where v_i is an assignment $V_i = x$ of a value to V_i and par_i is the assignment of values to its parents induced by the assignment $v = v_1 \cdots v_n$. In a *causally-interpreted* Bayesian net or *causal net*, the arrows in the DAG are interpreted as direct causal relations (Williamson, 2005) and the net can be used to infer the effects of interventions as well as to make probabilistic predictions (Pearl, 2000, Spirtes et al., 2000); in this case the Markov Condition is called the *Causal Markov Condition.*

4.3 Recursive Bayesian nets

A Recursive Bayesian Net is a Bayesian net defined over a finite set V of variables *whose values may themselves be RBNs.* A variable is called a *network variable* if one or more of its possible values is an RBN and a *simple variable* otherwise. A standard Bayesian net is an RBN whose variables are all simple.

An RBN x that occurs as the value of a network variable in RBN y is said to be at a *lower level* than y; the network variable in question is a *direct superior* of the variables in x, which are called its *direct inferiors.* Variables in the same net (i.e., at the same level) are *peers.* If an RBN contains no infinite descending chains—i.e., if each descending chain of nets terminates in a standard Bayesian net—then it is *well-founded.*[13] We restrict our attention to well-founded RBNs here.

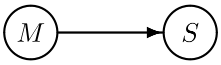
[12] Although Bayesian nets have been extended to handle certain continuous cases, we restrict attention to discrete variables in this paper.

[13] This corresponds to the notion of 'bottoming-out' in the mechanistic literature (see §2).

For simplicity of exposition, we shall also restrict our attention to the case in which each variable only occurs once in the RBN—in which case each variable has at most one direct superior. This will allow us to state our main points without having to digress by discussing questions to do with the consistency of an RBN and other technicalities (Williamson, 2005, §§10.4–10.5). Although this restriction is expedient, it is not essential: the theory of RBNs does admit variables with multiple occurrences.

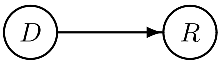
4.4 Example

To take a very simple example, consider an RBN on $V = \{M, S\}$, where M is the *DNA damage response mechanism* which takes two possible values, 0 and 1, while S is survival after 5 years which takes two possible values *yes* and *no*. The corresponding Bayesian net is:



$$P(M), P(S|M)$$

Suppose S is a simple variable but that M is a network variable, with each of its two values denoting a lower-level (standard) Bayesian network that represents a state of the DNA damage response mechanism. When M is assigned value 1 we have a net m_1 representing a functioning damage response mechanism, with a probabilistic dependence (and a causal connection) between damage D and response R:



$$P_{m_1}(D), P_{m_1}(R|D)$$

On the other hand, when M is assigned value 0 we have a net m_0 representing a malfunction of the damage response mechanism, with no dependence (and no causal connection) between damage D and response R:

$$P_{m_0}(D), P_{m_0}(R)$$

Since these two lower-level nets are standard Bayesian nets the RBN is well-founded and fully described by the three nets.

Note that, as this example shows, an RBN can be used to represent mechanisms in various states—in this case, the RBN represents a malfunctioning damage response mechanism as well as a functioning damage response mechanism.[14] It is possible to build an RBN representing just one of these mechanism states by taking the network variable to have a single possible value. Even if an RBN represents just one of a mechanism's states, it still models its hierarchical architecture.

4.5 Recursive Causal Markov Condition

If an RBN is to be used to model a mechanism, it is natural to interpret the arrows at the various levels of the RBN as signifying causal connections. Just as standard causally-interpreted Bayesian nets are subject to the Causal Markov Condition, a similar condition applies to causally-interpreted RBNs. This is called the *Recursive Causal Markov Condition*.

Let $\mathcal{V} = \{V_1, \dots, V_m\}$ $(m \geq n)$ be the set of variables of an RBN closed under the inferiority relation: i.e., $\mathcal{V}$ contains the variables in V, their direct inferiors, their direct inferiors, and so on. Then:

RCMC. Each variable in $\mathcal{V}$ is independent of those variables that are neither its effects (i.e., descendants) nor its inferiors, conditional on its direct causes (i.e., parents) and its direct superiors: $V_i \perp\!\!\!\perp \mathit{NID}_i \mid \mathit{DSup}_i \cup \mathit{Par}_i$ for each variable V_i, where NID_i is the set of non-inferiors-or-descendants of V_i and DSup_i is the set of direct superiors of V_i.

Note that, while some authors treat the Causal Markov Condition as a necessary truth, others argue against its universal validity (see, e.g., Williamson, 2005). We treat the Causal Markov Condition and RCMC as *modelling assumptions*, in need of testing or justification, rather than necessary truths.

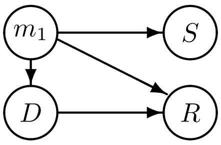
4.6 Inference

Inference in RBNs proceeds via a formal device called a *flattening*. Let $\mathcal{N} = \{V_{j_1}, \dots, V_{j_k}\} \subseteq \mathcal{V}$ be the network variables in $\mathcal{V}$. For each assignment $n = v_{j_1}, \dots, v_{j_k}$ of values to the network variables we can construct a standard Bayesian net, the *flattening* of the RBN with respect to n, denoted by $n^{\downarrow}$, by taking as nodes the simple variables in $\mathcal{V}$ plus the assignments $v_{j_1}, \dots, v_{j_k}$ to the network variables, and including an arrow from one variable to another if the former is a parent or direct superior of the latter in the original RBN. The conditional probability distributions are constrained by those in the original RBN: $P(V_i|\mathit{Par}_i \cup \mathit{DSup}_i) = P_{v_{j_i}}(V_i|\mathit{Par}_i)$ given in the RBN, where V_{j_i} is the direct superior of V_i. The Markov Condition holds in the flattening because the Recursive Causal Markov Condition holds in the RBN. (In the flattening, those arrows linking variables to their direct inferiors in the RBN would not

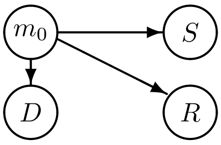
[14] The malfunctioning of mechanisms is of particular interest to, e.g., neuroscience (Craver, 2007, 124-125) and medicine (Nervi, 2010).

normally be interpretable causally, so the *Causal* Markov Condition is not satisfied.)[15]

In our example, for assignment m_1 of network variable M we have the flattening $m_1^{\downarrow}$:



with probability distributions $P(m_1) = 1$ and $P(S|m_1)$ determined by the top level of the RBN, and with $P(r_1|d_1 m_1) = P_{m_1}(r_1|d_1)$ determined by the lower level (similarly for d_0 and r_0). The flattening with respect to assignment m_0 is $m_0^{\downarrow}$:



Again $P(r_1|m_0) = P_{m_0}(r_1)$, etc. In each case the required conditional distributions are determined by the distributions given in the original RBN.

The flattenings suffice to determine a joint probability distribution over the variables in $\mathcal{V}$ via $P(v_1 \cdots v_m) = \prod_{i=1}^{m} P(v_i|par_i\, dsup_i)$ where the probabilities on the right-hand side are determined by a flattening induced by $v_1 \cdots v_m$. Having determined a joint distribution, the causally-interpreted RBN can be used to draw quantitative inferences in just the same way as can a standard causal net (Casini et al., 2011, §2).

Note that RBNs are more expressive than standard Bayesian nets. What can be encapsulated in a single RBN corresponds to the information in several standard Bayesian nets (the flattenings). In many cases the flattenings are mutually inconsistent, so cannot themselves be combined into a single standard Bayesian net.

4.7 Causal cycles

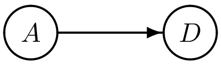
We are now in a position to see why causal cycles pose a conundrum for RBNs when used to model mechanisms. As we have seen, mechanisms with causal cycles are ubiquitous. Now, an RBN models causality at each level using directed *acyclic* graphs. It is important that the graph be acyclic because of the connection between RBNs and standard Bayesian nets: an RBN with a cyclic causal graph would lead to flattenings that themselves have cycles; these

[15] While these arrows would not normally be interpreted causally, the question arises as to whether they might be if Craver's views of mutual manipulability and causality are endorsed. See footnote 25.

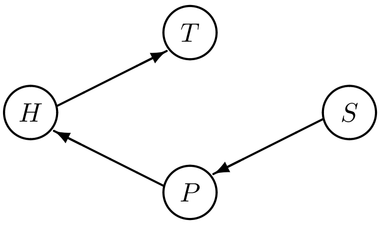
flattenings would fail to qualify as standard Bayesian nets and hence it would not be possible to define a joint distribution in the way described; therefore one would not be able to use the RBN for inference. Because causal cycles cannot be modelled directly in RBNs, it seems that RBNs cannot be suitable for modelling mechanisms after all.

Consider an example. Head injuries are often characterised by the following causal cycle (see §3): trauma causes swelling (oedema) which causes increased pressure on the brain (raised intracranial pressure) which causes oxygen deprivation (hypoxia) which in turn causes further trauma, and so on. Medical interventions to break this vicious circle include the use of the drug mannitol, which osmotically reduces swelling, and controlled hyperventilation, which reduces the partial pressure of carbon dioxide in the blood, which in turn produces cerebral vasoconstriction, reducing oedema. The important thing to note is that these interventions sever the connection between trauma and swelling.

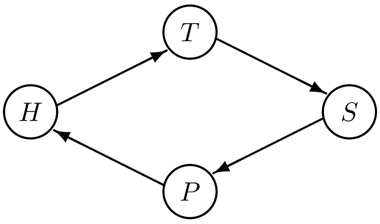
One might want to use an RBN to model this as follows. At the top level of the mechanism we might have two variables: action A taking possible values a_0 (no treatment), a_1 (treatment); and survival after 1 day D taking values d_0 (no), d_1 (yes). A is a cause of D at this level:



D might be considered a simple variable while A is modelled as a network variable. Value a_1 has causal graph:



where binary variables T, S, P, H stand for trauma, swelling, increased pressure and hypoxia respectively. a_0 has causal graph:



Clearly this last graph is not acyclic and hence the resulting hierarchical structure cannot be used as a basis for an RBN as standardly defined.

The question arises, then, whether the RBN formalism can be further developed in order to model mechanisms containing causal cycles.

5 Current approaches

The best way to find an answer to this question is to look at past attempts to handle causal cycles in ordinary, i.e. non-recursive, causal Bayesian nets. Two main types of solution can be distinguished, each of which may be used within the RBN framework, as we will show in §6.

5.1 *d*-separation in directed cyclic graphs

In the case of directed acyclic graphs (DAGs), there is an intuitive graphical criterion, called *d*-separation, that can be used to check which conditional independence relations are satisfied by any probability distribution that satisfies the Causal Markov Condition relative to a given graph G.

Definition 1 (*d*-separation) Let G be a DAG defined over the set of variables V. If $Q \subset V$ and $A, B \in V \setminus Q$, then A and B are d-separated given Q in G, in short $A \perp\!\!\!\perp_G B \mid Q$, iff there is no path U between A and B, such that

1. for every collider $\ldots \rightarrow C \leftarrow \ldots$ on U, $\mathit{Descendants}(C) \cap Q \neq \emptyset$,[16]
2. and no other vertex on U is in Q.

If $X \neq \emptyset, Y \neq \emptyset$ and Z are three disjoint sets, then X is d-separated from Y given Z iff every member of X is d-separated from every member of Y given Z. (cf. Spirtes et al., 2000, 44)

In the acyclic case, d-separation is equivalent to the Causal Markov Condition (Spirtes, 1995; Spirtes et al., 2000, 44): $X \perp\!\!\!\perp_G Y \mid Z$ iff $X \perp\!\!\!\perp_P Y \mid Z$, where $X, Y, Z \subset V$ and where the latter expression stands for 'X is probabilistically independent of Y given Z' for any probability distribution P that satisfies the Causal Markov Condition with respect to G.

This equivalence fails in directed cyclic graphs (DCGs), yet the following weaker implication holds as was shown by Pearl and Dechter (1996, 422, theorem 2).[17] Given a (possibly cyclic) graph and associated probability distribution, if $X \perp\!\!\!\perp_G Y \mid Z$, then $X \perp\!\!\!\perp_P Y \mid Z$, provided that (i) the variables in V all have a discrete and finite domain, (ii) the values of the variables of the system are uniquely determined by the disturbances, and (iii) the disturbances

[16] In this definition, $C \in \mathit{Descendants}(C)$ by convention.

[17] The paper by Pearl and Dechter (1996) extends previous results by Spirtes (1995) and Koster (1996) who have shown that the d-separation test is valid for cyclic graphs with linear equations and normal distributions over the error terms. Given that we restrict attention to discrete variables in this paper (see §4), we will only discuss Pearl and Dechter.

are uncorrelated.[18] In other words, even in the cyclic case the independencies induced by G can be read off directly by means of the d-separation criterion.[19]

Pearl and Dechter implicitly assume that the causal structure in question 'is stable' (1996, 421) and has reached 'equilibrium' (1996, 423)—cf. condition (ii) above: once the values of the disturbances are given, the values of all variables are uniquely determined. Their approach to the problem of cyclicity in ordinary Bayesian nets has a problem, however. As Neal (2000) has shown, their theorem is not true in general. He gives an example of a graph and an associated set of equations that satisfy the three conditions specified above, yet in which there are two variables that are d-separated by a third variable without being probabilistically independent conditional on that third variable. One possible way to salvage Pearl and Dechter's theorem is to require 'not only that [the disturbances] $U_1, \ldots, U_n$ uniquely determine [the endogenous variables] $X_1, \ldots, X_n$, but also that this unique solution for $X_1, \ldots, X_n$ can be obtained by a procedure in which the X_i are updated in accordance with the causal structure of the network. In such a casual [*sic*] dynamical procedure, each X_i is repeatedly replaced by the value computed for it from the corresponding U_i and the current values of its parents, according to the equation for that X_i, until a stable state is reached.' (Neal, 2000, 90)

One way to do this, is by making use of Dynamic Causal Bayesian nets.

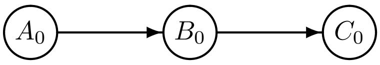
5.2 Dynamic Bayesian Nets

Dynamic Bayesian nets (DBNs) were developed in the late 1980s to model the change in a probability distribution over time (Dean and Kanazawa, 1989, §5). More recent developments can be found in Friedman et al. (1998), Ghahramani (1998), Murphy (2002), Bouchaffra (2010) and Doshi-Velez et al. (2011) for instance.

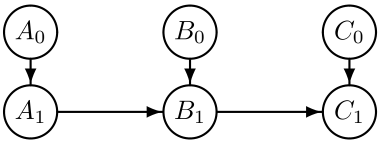
A DBN consists of two components. First, a *prior network* needs to be specified. This is a Bayesian network that is used to represent the probability distribution of the variables at the initial time 0. As an example, consider the probability distributions $P(A_0), P(B_0|A_0), P(C_0|B_0)$ together with the following graph:

[18] Disturbances are not explicitly mentioned in §4, but they can easily be introduced. Disturbances are variables that represent errors due to omitted factors (see Pearl, 2000, 27). The assumption of uncorrelated disturbances is not a severe restriction. Given a graph G and associated probability distribution P, such that not all disturbances are independent, one may construct an *augmented graph* G' in which all disturbances are independent. The augmented graph G' is obtained by adding, for each pair of dependent disturbances, a dummy root node as a common cause of the disturbances (Pearl and Dechter, 1996, 422).

[19] Unlike in the acyclic case, however, 'the joint distribution of feedback systems cannot be written as a product of the conditional distributions of each child variable, given its parents' (Pearl and Dechter, 1996, 420). Hence factorization, on which we relied in §4, cannot be applied to DCGs. This can be shown by means of a simple example by Spirtes et al. (2000, §12.1.2). Applying the factorization to the graph $X \leftrightarrows Y$, would lead to $P(X,Y) = P(X \mid Y) \times P(Y \mid X)$, which would mean that X and Y are independent, contrary to what the graph suggests.



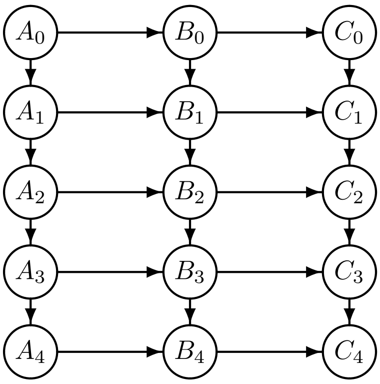
The second component of a DBN is a *transition network*, that can be used to represent the distribution of the variables at time 1 conditional on that at time 0. For example, we might have a transition network based on the following graph:



Here each variable at time 1 is directly influenced by its prior state as well as by its causes.

Note that the Markov Condition is an assumption underlying both these networks. It is also normally assumed that the process is *Markovian*, i.e., the variables at time $t + 1$ are probabilistically independent of those at times $0, \ldots, t - 1$ conditional on those at times t, and that the process is *stationary*, i.e., the distribution of variables at one time conditional on those at the previous time does not vary over time, so that the transition network remains valid for the transition from any time n to $n + 1$, not just for the transition from time 0 to time 1. (These assumptions can be relaxed, but obviously at a penalty of added computational complexity.)

For any particular time of interest, the DBN is *unrolled* by combining the prior network with sufficiently many copies of the transition network. For instance at time $t = 4$ we would need a network based on the following graph:



This unrolled network then determines a joint distribution over the variables from times 0 to 4.

The use of DBNs has been put forward as a way to handle causal cycles (see, e.g., Bernard and Hartemink, 2005). When there is a cycle linking two variables X and Y, that tends to be because X initially changes Y which later changes X, which in turn later changes Y and so on.[20] By time-indexing the variables in the cycle, one can *unwind* the cycle into a potentially infinite chain of the form $X_0 \longrightarrow Y_0 \longrightarrow X_1 \longrightarrow Y_1 \longrightarrow \cdots$. In general, a causal cycle can be unwound by time-indexing the variables and generating an acyclic DBN. This DBN can then be unrolled in the way described above.

6 Proposed solution

In this section we shall put forward a strategy for extending the RBN formalism to cope with causal cycles. This is a mixed strategy: some situations are handled one way, while other situations are handled another way. We distinguish between *static problems* and *dynamic problems*. A static problem is a situation in which a specific cycle reaches equilibrium—either due to negative feedback in the cycle or due to external factors—and where the equilibrium itself is of interest, rather than the process of reaching equilibrium. On the other hand a dynamic problem is a situation in which it is the change in the values of variables over time that is of interest: perhaps there is positive feedback, leading to a drift in the probability distribution of the variables in the cycle over time; perhaps there is negative feedback towards an equilibrium solution, but it is the progress of the cycle towards equilibrium that is of interest; perhaps the cycle variables oscillate between two or more distributions. Note that as to whether a problem is static or dynamic depends not only on the cycle in question but also on the interests of the modeller, as we suggested in §3. Our mixed strategy is then this: each static problem is tackled by appealing to the *d*-separation discussion of §5.1, while each dynamic problem is tackled by invoking the Dynamic Bayesian Net machinery introduced in §5.2.

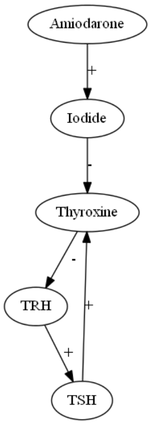
6.1 Static problems

In this case it is the equilibrium distribution itself that is of interest, rather than the values the variables take while reaching equilibrium. For each static problem within an RBN, we can attempt to model the probability distribution of the equilibrium solution. Our approach here will appeal to the use of *d*-separation in cyclic graphs, described in §5.1, to transform the cycle in question in the RBN into a Bayesian net that represents the corresponding equilibrium distribution, which we will call an *equilibrium network*.

Let us return to an earlier example of a stable cycle: the homeostatic thyroid cycle introduced in §3. This cycle, depicted in Figure 2, might well appear as a graph at some level in an RBN, perhaps with various malfunctioning variants appearing as other values of its direct superior. (Subclinical hypothyroidism, for example, involves an increase in TSH but normal levels of thyroxine, while primary hypothyroidism involves high levels of TSH and low

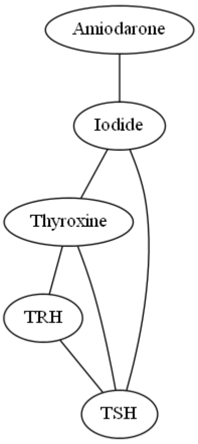
[20] Simultaneous causation and backwards causation do not fit this picture. However, such cases rarely if ever occur in models of mechanisms, so we set them aside in this paper.

levels of thyroxine.) We will consider a slightly augmented example, consisting of the thyroid cycle together with the following process. Amiodarone is a drug, commonly used to treat cardiac arrhythmias, that contains lots of iodine (37% by weight). One common adverse effect of this drug is hypothyroidism, which is an abnormal reduction in the concentration of thyroxine in the blood. This occurs because the iodine contained in amiodarone causes a reduction in the rate of iodide oxidation by the thyroid by a mechanism known as the Wolff-Chaikoff effect. The following should be interpreted as a directed cyclic graph with *Amiodarone*, *Iodide*, etc. as vertices representing random variables, and not just as a schematic picture as was the case in section §3):



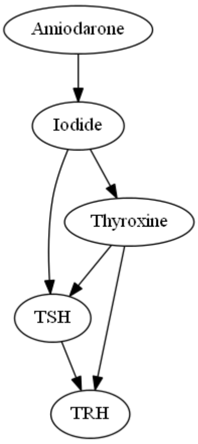
The procedure is first to transform this graph into an undirected *moral graph*, which is formed by 'marrying' the parents of each variable in the previous cyclic graph by adding an edge between them, and then adding an edge between each pair of variables that are connected by an arrow in the cyclic graph:[21]

[21] This procedure is outlined by Lauritzen et al. (1990). Their alternative test for *d*-separation is equivalent to the one specified in §5.1 and is used by Pearl and Dechter (1996) in their discussion of *d*-separation for directed cyclic graphs.



The key point here is that separation in the moral graph is equivalent to *d*-separation in the directed cyclic graph, which, as we saw in §5.1, implies conditional probabilistic dependence.[22] Thus for example amiodarone is probabilistically independent of thyroxine conditional on iodide level.

This undirected graph can then be transformed into a directed acyclic graph that satisfies the Markov Condition with respect to the underlying probability distribution (see the Appendix for an algorithm):



Note that the arrows in this graph are not to be interpreted causally. This equilibrium network is merely a formal device for representing a joint distribution.

Finally, the resulting graph can be substituted for the corresponding cyclic graph in the original RBN, and, by specifying the probability distribution of

[22] Separation in undirected graphs (such as moral graphs) is defined as follows: let G be an undirected graph with vertex set V, then two sets of vertices $X, Y \subseteq V$ are separated by $Z \subseteq V$ if and only if every path (sequence of undirected edges) from each vertex in X to each vertex in Y contains some vertex in Z.

each variable conditional on its parents, the network can be used for inference as detailed in §4.

6.2 Dynamic problems

In this situation it is the change in the values of variables over time that is of interest. The approach we take in this dynamic case is to unwind the cycles by time-indexing the variables, and apply the DBN formalism to represent the probability distribution of the variables in the cycle as it evolves over time.[23]

Let us return to our head trauma example. We gave an RBN representation of the relevant mechanisms including a_0 which has the causal graph depicted in figure 7:

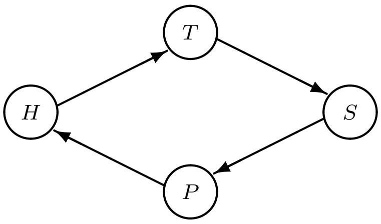


Fig. 7 Cyclic graph

This is a cyclic graph and needs to be transformed into a directed acyclic graph if we are to apply the RBN inference machinery. The key point to note is that the causal cycle is not instantaneous. A change in H changes T slightly later, which changes S slightly later, changing P later still, followed by a subsequent change in H, and so on. The point is that it is not the initial change in H that leads via a causal cycle to a change in itself, but rather that the initial change in H leads via a causal cycle to a change in a later value of H. Indexing the variables by time makes this temporal aspect explicit.

Recall that a DBN consists of two components. First, a *prior network* needs to be specified. This is a Bayesian network that is used to represent the probability distribution of the variables at time 0. The following graph, for instance, can be used in the prior network (though computationally it may not be the most convenient one, see below):

[23] *Prima facie*, this approach runs counter to Bechtel's appeal not to model mechanisms sequentially (see §2). By means of the transition network, however, the cyclic organization is captured as well. We would like to thank Michael Wilde for pointing out this seeming incongruity.

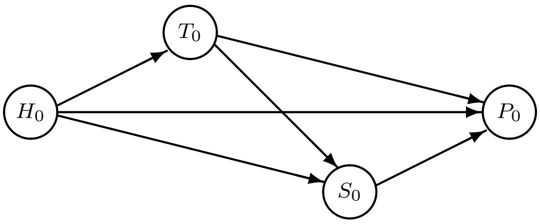


Fig. 8 Complete graph for prior network

This is a complete graph (there is an arrow between each pair of nodes) so the Markov Condition is trivially satisfied. But a complete graph can be computationally demanding: as the number of nodes in a complete graph increases, there is an exponential increase in both the number of probabilities that need to be specified in the corresponding Bayesian net and in the time it takes to perform inferences using the Bayesian net. Thus it is desirable to use a sparser graph if possible. In the static case, a sparser graph was obtained by generating a DAG via the moral graph. This method is not recommended in the dynamic case because, as we saw in §5.1, it only appears to be guaranteed to work in equilibrium situations. Instead we recommend a hypothesise and test methodology for obtaining a sparser graph. First, a DAG can be hypothesised by unwinding the cycle in figure 7 and considering the causal connections between the variables at the initial time 0 (see figure 9):

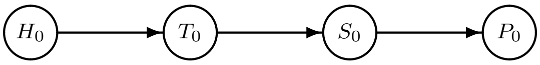


Fig. 9 Sparse graph for prior network

In general, the Markov condition may fail in a causal graph that has been constructed in this way from a cyclic causal graph.[24] Second, therefore, one needs to test whether the Markov Condition is satisfied to within some specified tolerance level. If not, one should add further arrows until the Markov Condition is satisfied. See Williamson (2005, §3.6) for an algorithm for prioritising which arrows to add in order to more closely satisfy the Markov

[24] To take a concrete example, suppose that variables A and B directly cause C which directly causes D which in turn directly causes each of A and B. A causal DAG at time 0 obtained by unwinding this cycle might have arrows from A_0 and B_0 to C_0 and an arrow from C_0 to D_0. The Markov condition requires that A_0 and B_0 be probabilistically independent. But these two variables have a common cause not represented in the graph—the previous instance of D—which can render them probabilistically dependent. Thus the Markov condition can fail in this causal graph.

Condition. As in the static case, the resulting DAG is not to be causally interpreted: it is merely a formal device for representing a probability distribution. The form of the resulting DAG depends on the underlying probability distribution, but it will contain the initial causal DAG as a subgraph. We shall suppose for simplicity of exposition that in our example the prior network is based on the causal DAG depicted in figure 9—i.e., that no further arrows need to be added to satisfy the Markov condition.

Having specified a prior network, we need to specify a *transition network* that can be used to represent the distribution of the variables at time 1 conditional on that at time 0, such as that based on the following graph:

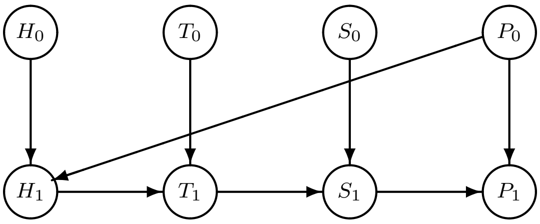


Fig. 10 Graph for transition network

This graph is produced by taking as nodes the variables at times 0 and 1, and adding arrows to each variable at time 1 from its prior state and its causes.

One can then unroll the network by combining the prior network (figure 9) with sufficiently many copies of the transition network (figure 10). For instance at time $t = 4$ we would need a network based on the following graph:

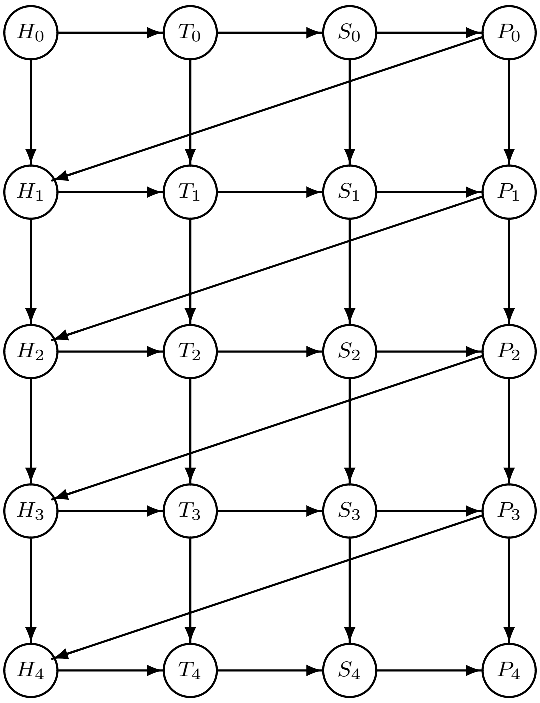


Fig. 11 Graph for unrolled network

This network can then be substituted for the corresponding causal cycle in an RBN. The joint distribution over all the variables in the RBN can be calculated by forming the flattenings as usual.

Thus in the dynamic case we have a two-step process: the cycle is first *unwound* (and, if necessary, further arrows are added to ensure that the Markov condition holds) to form a DBN, which in turn is *unrolled* into a standard Bayesian net representing the distribution of the variables up to a certain time.

6.3 Specifying the probability distribution of an RBN with cycles

To sum up, our approach involves replacing causal cycles by directed acyclic graphs in the RBN. The kind of replacement depends on whether or not the temporal dimension is of interest or is required: if so, we apply the DBN formalism, if not, we apply the extension of *d*-separation methods to cyclic graphs.

In §4 an RBN was characterised as a Bayesian net defined over a finite set V of variables whose values may themselves be RBNs. We saw that this characterisation is inadequate, given the use of RBNs to model mechanisms, because causal cycles are pervasive in mechanisms, while the graph of a Bayesian net must be acyclic. We therefore need to generalise the notion of an RBN so that the RBN itself is permitted to contain cycles. This allows one to retain a causal interpretation of an RBN: in a causally interpreted RBN, each arrow in the RBN is interpreted as a direct causal connection. For each cycle in the RBN we also need to provide further information, in order that the RBN can determine a joint probability distribution and thereby be used for quantitative inference. If there is no equilibrium state of the variables in the cycle then we have a dynamic problem, so we must specify a prior Bayesian network and transition network involving the peers of the variables in the cycle. If there is an equilibrium state, then the particular application will determine whether a dynamic approach or a static approach is required. In this case one may need to specify an equilibrium network, i.e., a Bayesian network representing the equilibrium state of the variables in and around the cycle, instead of, or as well as, a prior network and a transition network.

In Williamson (2005, Chapter 10) it was suggested that, in cases where a probability distribution is constrained—rather than uniquely determined—by available quantitative information, one should use maximum entropy methods to determine a particular probability distribution that satisfies those constraints. Of course there is some controversy concerning maximum entropy methods (see, e.g., Seidenfeld, 1987; Williamson, 2010). But if this route is accepted, then the prior, transition and equilibrium networks can be generated as needed, on the fly, from the constraints imposed by the probability distribution of each variable conditional on its parents given in the (possibly cyclic) RBN: the graphs in the networks can be constructed as outlined above, while the probability distribution of each variable conditional on its parents in the graph can be chosen to be the distribution, from all those that satisfy constraints specified in the RBN, that has maximum entropy.

We should note that it is common to distinguish single-case models from generic models. The former kind of model represents a particular case while the latter is repeatedly instantiatable. An RBN is a *generic* model of a mechanism: it can be instantiated in a variety of single cases. But often it is only in the context of a particular instantiation that one can determine whether one is tackling a static or a dynamic problem. This is because, at least in contingent cycles, as to whether there is an equilibrium state or not can depend on the particular case. Moreover, the particular problem can influence whether one is concerned with the progress towards equilibrium or the equilibrium itself. Hence an RBN can be viewed as a schematic representation of a mechanism, with the details to be filled in according to the application in question.

Note too that arrows in the RBN model of a mechanism are all causally interpreted, but when the above strategy for handling cycles is executed, arrows in the resulting network may no longer all be causal. Thus when using a cyclic RBN to predict the effects of an intervention, for instance, one must first perform the intervention on the cyclic RBN (deleting arrows into the node which is set by the intervention; Pearl, 2000, pp. 22–23), and only then apply the strategy for handling cycles. Finally, the inference methods of §4.6 may be applied, requiring further transformations in order to produce the flattenings.

These two points reinforce the claim that it is the (possibly cyclic) RBN that is the fundamental model of the mechanism, with transformations of this section and §4.6 to be applied only when required for inference in particular applications.

6.4 Related work

As far as we are aware, there is only one other attempt to use hierarchical versions of Bayesian nets to model mechanisms for the purpose of quantitative inference. Gebharter and Kaiser (2012) do not adopt the RBN framework but rather represent mechanisms by a hierarchy of disjoint causal Bayesian networks. This has its advantages and its disadvantages over our RBN approach. On the one hand they do not need a modelling assumption such as RCMC to tie the levels of the mechanism together, so their assumptions are weaker. On the other hand, it is not possible under their approach to represent the joint distribution over all the variables in the hierarchy; this results in a narrower range of inferences that can be drawn from their representation. For instance, one cannot use the value of a variable at one level of the hierarchy to help predict the value of a variable at another level, in the absence of a single causal network that includes both variables. The authors recognise the importance of handling cycles appropriately in order to model mechanisms, and they recommend a time-indexing approach similar to that which we advocate in the case of dynamic problems. We would argue that this is not the appropriate strategy in the case of static problems, because it introduces irrelevant details into the representation and because it can lead to unnecessary computational cost.

7 Summary and concluding remarks

In this paper, we have presented one possible quantitative approach to modelling mechanisms, which makes use of Recursive Bayesian nets. Causal cycles, if present in the RBN, are replaced by directed acyclic graphs in order to perform inference using inference techniques for standard Bayesian nets. The kind of replacement depends on whether or not the temporal dimension is of interest or is required: if so, we apply the DBN formalism, if not, we apply the extension of *d*-separation methods to cyclic graphs.

To end this paper, we would like to suggest some questions for further research. First, it would be interesting to further explore the consequences of our approach for philosophy of science. For instance, we hypothesise that the RBN framework may shed further light on Craver's mutual manipulability account of constitutive relevance (and vice versa). Interventions on causal Bayesian nets have been discussed extensively (see Pearl, 2000 and Spirtes et al., 2000).

These notions carry over, *mutatis mutandis*, to RBNs, and may help to analyse Craver's notion of 'interlevel intervention' and the interlevel experiments on which it is based.[25]

Second, the merits and disadvantages of our approach as compared to other quantitative accounts of mechanistic modelling should be further explored. In §6.4 we briefly discussed the work of Gebharter and Kaiser (2012). Yet the precise relation between our account and, for example, Bechtel's dynamic systems analysis (mentioned in §2) remains an open question.

Third, RBNs open up the possibility of algorithms for mechanism discovery. The advent of causal Bayesian nets led to a range of algorithms for causal discovery (see, e.g., Spirtes et al., 2000, 2010). Because of the close relation between RBNs and Bayesian nets, it is plausible that algorithms for learning causal structure might be extended to algorithms for learning causal and mechanistic structure simultaneously. The main task would be to distinguish between causal and superiority (i.e., mechanistic hierarchy) relations. The difference between causal manipulation and mutual manipulability might offer a starting point in this respect. If progress can be made here, it could have enormous payoffs for those—such as bioinformatics researchers and pharmaceutical companies—who are engaged in 'closing the inductive loop' by automating both scientific experimentation and scientific discovery.

Finally, the limits of our approach should be further explored. On the non-formal side, we fully acknowledge that our framework leaves out some interesting features of mechanisms that are captured by alternative ways of representing them. For example, a large part of the functioning of a mechanism depends on the spatial organization of its lower-level components, yet neither causal Bayesian nets nor Recursive Bayesian Nets offer a natural way of representing this spatial organization. Likewise, mechanistic diagrams such as those presented in §3 are often easier to grasp and to (humanly) reason with than ordinary or Recursive Bayesian Nets. (See Perini, 2005a,b,c for interesting discussions of the role of diagrams and visual representations in scientific reasoning; see also, among many others, Craver, 2006 and Bechtel and Abrahamsen, 2005 for a discussion of diagrams in mechanistic contexts.)

On the formal side, we treat the Causal Markov Condition and the Recursive Causal Markov Condition as modelling assumptions rather than necessary truths. Whereas the limits of the former have been discussed extensively in the literature,[26] those of the latter remain to be inspected in detail.

Also on the formal side, the question arises as to how the framework presented here can be extended to handle certain continuous cases. Modelling continuous cases with cyclic causality by means of discrete variables may lead to problems; for example, spurious instabilities may arise in the model even

[25] For a detailed account of mutual manipulability, see Craver (2007, 152–160). For a recent critical discussion of Craver's claim that interlevel constitutive relations cannot be causal, and whether this claim is compatible with his mutual manipulability account of constitutive relevance, see Leuridan (2012).

[26] See, among others, Hausman and Woodward (1999, 2004a,b), Cartwright (2001, 2002), Williamson (2005) and Steel (2006).

when the original system itself is stable (see Pearl and Dechter, 1996, 425). Yet we do not expect any major difficulties here. Causal Bayesian nets can easily be defined over continuous variables (see footnote 12) and in fact, the problem of automated causal discovery is often easier in the continuous case (e.g., assuming normally distributed variables) than in the discrete case. Likewise, RBNs can easily be defined over continuous variables as well. Our choice to restrict ourselves to the discrete case was motivated by our endeavour to limit technicalities as far as possible.

Appendix: Transforming a moral graph into a directed acyclic graph

Here we present the algorithm of Williamson (2005, §5.7) for transforming an undirected graph $\mathcal{G}$ into a directed acyclic graph $\mathcal{H}$ which preserves the required independencies (Williamson, 2005, Theorem 5.3): if Z d-separates X from Y in the directed acyclic graph $\mathcal{H}$ then X and Y are separated by Z in the undirected graph $\mathcal{G}$; this separation in $\mathcal{G}$ implies that X and Y are probabilistically independent conditional on Z; hence, d-separation in $\mathcal{H}$ implies that X and Y are probabilistically independent conditional on Z. Thus $\mathcal{H}$ can be used as the graph of a Bayesian network.

An undirected graph is *triangulated* if for every cycle involving four or more vertices there is an edge in the graph between two vertices that are non-adjacent in the cycle. The first step of the procedure is to construct a triangulated graph $\mathcal{G}^T$ from the undirected graph $\mathcal{G}$. One of a number of standard triangulation algorithms can be applied to construct $\mathcal{G}^T$ (see, e.g., Neapolitan, 1990, §3.2.3; Cowell et al., 1999, §4.4.1).

Next, re-order the variables in V according to *maximum cardinality search* with respect to $\mathcal{G}^T$: choose an arbitrary vertex as V_1; at each step select the vertex which is adjacent to the largest number of previously numbered vertices, breaking ties arbitrarily. Let $D_1, \ldots, D_l$ be the cliques (i.e., maximal complete subgraphs) of $\mathcal{G}^T$, ordered according to highest labelled vertex. Let $E_j = D_j \cap (\bigcup_{i=1}^{j-1} D_i)$ and $F_j = D_j \backslash E_j$, for $j = 1, \ldots, l$.

Finally, construct a directed acyclic graph $\mathcal{H}$ as follows. Take variables in V as vertices. Step 1: add an arrow from each vertex in E_j to each vertex in F_j, for $j = 1, \ldots, l$. Step 2: add further arrows to ensure that there is an arrow between each pair of vertices in $D_j, j = 1, \ldots, l$, taking care that no cycles are introduced (there is always some orientation of an added arrow which will not yield a cycle).

Acknowledgements

We would like to thank Lorenzo Casini, George Darby, Phyllis Illari, Mike Joffe, Federica Russo, and Michael Wilde for their helpful comments on this paper. Jon Williamson's research is supported by the UK Arts and Humanities Research Council. Bert Leuridan is Postdoctoral Fellow of the Research Foundation—Flanders (FWO).

References

Bechtel, W. (2011). Mechanism and biological explanation. *Philosophy of Science*, 78(4):533–557.

Bechtel, W. and Abrahamsen, A. (2005). Explanation: a mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2):421–441.

Bernard, A. and Hartemink, A. (2005). Informative structure priors: Joint learning of dynamic regulatory networks from multiple types of data. In Altman, R., Dunker, A. K., Hunter, L., Jung, T., and Klein, T., editors, *Proceedings of the Pacific Symposium on Biocomputing (PSB05)*, pages 459–470, Hackensack, NJ. World Scientific.

Boon, N. A., Colledge, N. R., Walker, B. R., and Hunter, J. A., editors (2006). *Davidson's Principles & Practice of Medicine*. Churchill Livingstone, Edinburgh, 20th edition.

Bouchaffra, D. (2010). Topological dynamic bayesian networks. In *Proceedings of the Twentieth International Conference on Pattern Recognition*, pages 898–901. IEEE.

Cartwright, N. (2001). What is wrong with Bayes nets? *The Monist*, 84(2):242–264.

Cartwright, N. (2002). Against modularity, the causal Markov condition, and any link between the two: Comments on Hausman and Woodward. *The British Journal for the Philosophy of Science*, 53(3):411–453.

Casini, L., Illari, P. M., Russo, F., and Williamson, J. (2011). Models for prediction, explanation and control: recursive Bayesian networks. *Theoria*, 26(1):5–33.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. Springer-Verlag, Berlin.

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, 153(3):355–376.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Clarendon Press, Oxford.

Crunelli, V. and Hughes, S. W. (2010). The slow (<1 Hz) rhythm of non-REM sleep: a dialogue between three cardinal oscillators. *Nature Neurosciences*, 13(1):9–17.

Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, 5(3):142–150.

Doshi-Velez, F., Wingate, D., Tenenbaum, J., and Roy, N. (2011). Infinite dynamic bayesian networks. In Getoor, L. and Scheffer, T., editors, *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 913–920. Omnipress.

Friedman, N., Murphy, K. P., and Russell, S. J. (1998). Learning the structure of dynamic probabilistic networks. In Cooper, G. F. and Moral, S., editors, *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 139–147, San Mateo, CA. Morgan Kaufmann.

Gebharter, A. and Kaiser, M. I. (2012). Causal graphs and mechanisms. In Hüttemann, A., Kaiser, M. I., and Scholz, O., editors, *Explanation in the Special Sciences. The Case of Biology and History*, Synthese Library. Springer, Dordrecht.

Ghahramani, Z. (1998). Learning dynamic bayesian networks. In Giles, C. and Gori, M., editors, *Adaptive Processing of Sequences and Data Structures*, volume 1387 of *Lecture Notes in Computer Science*, pages 168–197. Springer Berlin / Heidelberg. 10.1007/BFb0053999.

Glennan, S. S. (1996). Mechanisms and the nature of causation. *Erkenntnis*, 44(1):49–71.

Glennan, S. S. (2002). Rethinking mechanistic explanation. *Philosophy of Science*, 69(3):S342–S353.

Glennan, S. S. (2005). Modeling mechanisms. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2):443–464.

Gottlieb, D. J., Yenokyan, G., Newman, A. B., O'Connor, G. T., Punjabi, N. M., Quan, S. F., Redline, S., Resnick, H. E., Tong, E. K., Diener-West, M., and Shahar, E. (2010). Prospective study of obstructive sleep apnea and incident coronary heart disease and heart failure. *Circulation*, 122(4):352–360.

Grandner, M., Hale, L., Moore, M., and Patel, N. V. (2010). Mortality associated with short sleep duration: The evidence, the possible mechanisms, and the future. *Sleep Medicine Reviews*, 14(3):191–203.

Hausman, D. M. and Woodward, J. (1999). Independence, invariance and the causal Markov condition. *The British Journal for the Philosophy of Science*, 50(4):521–583.

Hausman, D. M. and Woodward, J. (2004a). Manipulation and the causal Markov condition. *Philosophy of Science*, 55(5):147–161.

Hausman, D. M. and Woodward, J. (2004b). Modularity and the causal Markov condition: a restatement. *The British Journal for the Philosophy of Science*, 55(1):147–161.

Koster, J. T. A. (1996). Markov properties of nonrecursive causal models. *Annals of Statistics*, 24(5):2148–2177.

Lauritzen, S., Dawid, A., Larsen, B., and Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, 20(5):491–505.

Lazebnik, Y. (2002). Can a biologist fix a radio?—or, what I learned while studying apoptosis. *Cancer Cell*, 2:179–182.

Leuridan, B. (2010). Can mechanisms really replace laws of nature? *Philosophy of Science*, 77(3):317–340.

Leuridan, B. (2012). Three problems for the mutual manipulability account of constitutive relevance in mechanisms. *The British Journal for the Philosophy of Science*, 63(2):399–427.

Leuridan, B. (2014). The structure of scientific theories, explanation, and unification. A causal-structural account. *The British Journal for the Philosophy of Science*. forthcoming.

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1):1–25.

Marshall, L., Helgadóttir, H., Mölle, M., and Born, J. (2006). Boosting slow oscilllations during sleep potentiates memory. *Nature*, 444(7119):610–613.

McNicholas, W. and Bonsignore, M. (2007). Sleep apnoea as an independent risk factor for cardiovascular disease: current evidence, basic mechanisms and research priorities. *European Respiratory Journal*, 29(1):156–178.

Mitchell, S. (2009). *Unsimple Truths: Science, Complexity, and Policy.* University of Chicago Press, Chicago, IL.

Murphy, K. P. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning.* PhD thesis, Computer Science, University of California, Berkeley.

Neal, R. (2000). On deducing conditional independence from *d*-separation in causal graphs with feedback: The uniqueness condition is not suffient. *Journal of Artificial Intelligence Research*, 12:87–91.

Neapolitan, R. E. (1990). *Probabilistic reasoning in expert systems: theory and algorithms.* Wiley, New York.

Nervi, M. (2010). Mechanisms, malfunctions and explanation in medicine. *Biology and Philosophy*, 25(2):215–228.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge.

Pearl, J. and Dechter, R. (1996). Identifying independencies in causal graphs with feedback. In *In Uncertainty in Artificial Intelligence: Proceedings of the Twelfth Conference*, pages 420–426, San Mateo, CA. Morgan Kaufmann.

Perini, L. (2005a). Explanation in two dimensions: Diagrams and biological explanation. *Biology and Philosophy*, 20(2-3):257–269.

Perini, L. (2005b). The truth in pictures. *Philosophy of Science*, 72(1):262–285.

Perini, L. (2005c). Visual representations and confirmation. *Philosophy of Science*, 72(5):913–926.

Seidenfeld, T. (1987). Entropy and uncertainty. In MacNeill, I. B. and Umphrey, G. J., editors, *Foundations of Statistical Inference*, pages 259–287. D. Reidel, Dordrecht.

Spirtes, P. (1995). Directed cyclic graphical representation of feedback models. In Besnard, P. and Hanks, S., editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 491–498. Morgan Kaufmann, San Mateo, CA.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search.* MIT Press, Cambridge, MA.

Spirtes, P., Glymour, C., Scheines, R., and Tillman, R. (2010). Automated search for causal relations—theory and practice. In Dechter, R., Geffner, H., and Halpern, J. Y., editors, *Heuristics, probability and causality: a tribute to Judea Pearl*, pages 467–506. College Publications, London.

Steel, D. (2006). Comment on Hausman & Woodward on the causal Markov condition. *British Journal for the Philosophy of Science*, 57(1):219–231.

Williamson, J. (2005). *Bayesian nets and causality: philosophical and computational foundations.* Oxford University Press, Oxford.

Williamson, J. (2010). *In defence of objective Bayesianism*. Oxford University Press, Oxford.

Woodward, J. (2002). What is a mechanism? A counterfactual account. *Philosophy of Science*, 69(3):S366–S377.

Woodward, J. (2003). *Making Things Happen. A Theory of Causal Explanation*. Oxford University Press, New York.

Yumino, D., Redolfi, S., Ruttanaumpawan, P., Su, M., Smith, S., Newton, G. E., Mak, S., and Bradley, T. D. (2010). Nocturnal rostral fluid shift. *Circulation*, 121(14):1598 –1605.