RECURSIVE BAYESIAN NETS FOR PREDICTION, EXPLANATION AND CONTROL IN CANCER SCIENCE

A Position Paper

Lorenzo Casini, Phyllis McKay Illari, Federica Russo, Jon Williamson
Philosophy, University of Kent, Canterbury, UK
l.casini@kent.ac.uk, p.mckay@kent.ac.uk, f.russo@kent.ac.uk, j.williamson@kent.ac.uk

Keywords: Bayesian network, Recursive Bayesian network, Prediction, Explanation, Control, Mechanism, Causation, Causality, Cancer, DNA Damage Response

Abstract: The Recursive Bayesian Net formalism was originally developed for modelling nested causal relationships. In this paper we argue that the formalism can also be applied to modelling the hierarchical structure of physical mechanisms. The resulting network contains quantitative information about probabilities, as well as qualitative information about mechanistic structure and causal relations. Since information about probabilities, mechanisms and causal relations are vital for prediction, explanation and control respectively, a recursive Bayesian net can be applied to all these tasks.
We show how a Recursive Bayesian Net can be used to model mechanisms in cancer science. The highest level of the proposed model will contain variables at the clinical level, while a middle level will map the structure of the DNA damage response mechanism and the lowest level will contain information about gene expression.

1 INTRODUCTION

Bayesian networks were originally developed to model probabilistic and causal relationships (Pearl, 1988). In the last two decades Bayesian nets have become the model of choice for making quantitative predictions and for deciding which variables to intervene on in order to control variables of interest. Thus a Bayesian net can be used to answer questions such as: *given that patient has has treatment t what is the probability $P(r|t)$ that their cancer will recur in the next* 5 *years?* and: *on which variables should we intervene in order to minimise the probability of recurrence?* Causal information is important here because it is only worth intervening on the *causes* of recurrence, not on other variables which might be indicators or evidence of recurrence.

The causal structure modelled by a Bayesian net can also help answer certain simple explanatory questions, such as: *what was the chain of events that led up to the recurrence of the patient's cancer?* But often we want to be able to offer explanations, not in this backwards, ætiological sense, but in a downwards, mechanistic sense. In order to answer *how did the patient's cancer recur?* we may need to specify the lower-level activities of the relevant cancer mechanism and the corresponding cancer response mechanisms. To answer such explanatory questions a model needs to represent the relevant mechanisms, including their hierarchical organisation.

Bayesian nets have been extended to model hierarchy in a number of ways. For example, *recursive Bayesian multinets* model context-specific independence relationships and decisions (Peña et al., 2002), *recursive relational Bayesian networks* model relational structure and more complex dependence relationships (Jaeger, 2001), *object-oriented Bayesian networks* can simplify the structure of large and complex Bayesian nets (Koller and Pfeffer, 1997), *hierarchical Bayesian networks* offer a very general means of modelling arbitrary lower-level structure (Gyftodimos and Flach, 2002) and *recursive Bayesian networks* were developed to model nested causal relationships (Williamson and Gabbay, 2005). In this paper we shall show how recursive Bayesian networks can also be used to model mechanisms, thus providing an integrated modelling formalism for prediction, explanation and control. This is important from the

AI perspective of providing models that can be used to answer a variety of queries in decision support systems, but also from the bioinformatics perspective of providing models that can integrate a variety of data sources with the more qualitative knowledge of the basic science involved.

In §2 we introduce the recursive Bayesian network formalism and show how it can be used to model mechanisms. In §3 we explain the current understanding of mechanisms in the philosophical literature, introduce the relevant cancer science mechanisms and show how the cancer science mechanisms fit the philosophical characterisation of mechanisms. In §4 we introduce the variables under consideration and the data sources to be used to construct the model. We summarise and outline future research in §5.

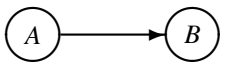
2 RECURSIVE BAYESIAN NETS

Recursive Bayesian networks (RBNs) were originally developed in (Williamson and Gabbay, 2005) to model nested causal relationships such as *[smoking causing cancer] causes tobacco advertising restrictions which prevent smoking which is a cause of cancer*. In this section we introduce the RBN formalism in the context of modelling mechanisms rather than nested causality.

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. These are linked by the *Markov Condition* which says that each variable is probabilistically independent of its non-descendants, conditional on its parents, written $V_i \perp\!\!\!\perp ND_i \mid Par_i$. 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 make probabilistic predictions (Pearl, 2000); in this case the Markov Condition is called the *Causal Markov Condition*.

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 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; variables in Y are the *direct superiors* of variables in X while variables in the same net are *peers*. If an RBN contains no infinite descending chains—i.e., if each decending chain of nets terminates in a standard Bayesian net—then it is *well-founded*. We restrict our attention to well-founded RBNs here.

To take a very simple example, consider an RBN on $V = \{A, B\}$, where A is *kind of tumour* which takes two possible values 0 and 1 while B is survival after 5 years which takes two possible values *yes* and *no*. The corresponding Bayesian net is:

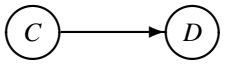


$$P(A), P(B|A)$$

Suppose B is a simple variable but that A is a network variable, with each of its two values denoting a lower-level (standard) Bayesian network that represents connections between gene expression levels in the corresponding kind of tumour. When A is assigned value 0 we have a net a_0 with no dependence between gene expression levels C and D:

$$P_{a_0}(C), P_{a_0}(D)$$

On the other hand, under assignment a_1, C and D are dependent:



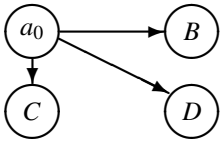
$$P_{a_1}(C), P_{a_1}(D|C)$$

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

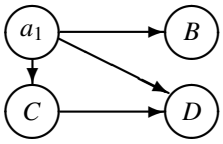
Since an RBN is a Bayesian net, the Markov Condition is imposed. But RBNs are subject to a further condition, the *Recursive Markov Condition*, which says that each variable is probabilistically independent of those variables that are neither its inferiors nor peers, conditional on its direct superiors, written $V_i \perp\!\!\!\perp NIP_i \mid DSup_i$. Let $\mathcal{V} = \{V_1, \ldots, V_m\}$ $(m \geq n)$ 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. Let $\mathcal{N} = \{V_{i_1}, \ldots, V_{i_k}\} \subseteq \mathcal{V}$ be the network variables in $\mathcal{V}$. For each assignment $n = v_{i_1}, \ldots, v_{i_k}$ 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 $v_{i_1}, \ldots, v_{i_k}$ to the network variables (these can be thought of as variables that can only take one possible value, i.e., constants), 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: $P(V_i|Par_i \cup DSup_i)$ must match the $P_{v_{i_j}}(V_i|Par_i)$ given in the RBN. If each variable has at most one direct superior in the RBN then this will uniquely determine the required distribution $P(V_i|Par_i \cup DSup_i)$; in other cases we follow (Williamson and Gabbay, 2005, §5) and take the distribution to be that, from all those that satisfy the constraints, which has maximum entropy. The Markov Condition holds in the flattening because the Markov Condition and Recursive Markov Condition hold in the RBN.

In our example, for assignment a_0 of network variable A we have the flattening $a_0^{\downarrow}$:



with probability distributions $P(a_0) = 1, P(B|a_0)$ determined by the top level of the RBN and with $P(d_1|a_0) = P_{a_0}(d_1)$ and similarly for d_0, c_0 and c_1. The flattening with respect to assignment a_1 is:



Again $P(d_1|c_1a_1) = P_{a_1}(d_1|c_1)$ etc. In each case the required conditional distributions are fully determined by the distributions given in the original RBN.

As long as certain consistency requirements are satisfied (Williamson and Gabbay, 2005, §4), 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 $v_1 \cdots v_m$.

With a joint distribution the model can be used for prediction. For example, the probability that D is expressed at level 1 and that the patient will survive 5 years is $P(b_1d_1) = P(a_0b_1d_1) + P(a_1b_1d_1) = P(b_1|a_0)P(a_0)P_{a_0}(d_1) + P(b_1|a_1)P(a_1)(P_{a_1}(d_1|c_1)P_{a_1}(c_1) + P_{a_1}(d_1|c_0)P_{a_1}(c_0))$. More than that, if at each level the arrows in the RBN can be interpreted causally then the model can be used for control and ætiological explanation: one might cite cancer type 0 as the reason a patient survived 5 years. If the inter-level relations match that of mechanistic composition then the model can be used for mechanistic explanation. Thus the gene expression levels $C = 0$ and $D = 1$ and the link between the two might explain survival. And one can use an RBN to reason across levels: by intervening on expression level D one might increase the probability of survival.

3 MECHANISMS IN PHILOSOPHY OF SCIENCE AND IN CANCER SCIENCE

The main thesis of this paper is that recursive Bayesian are legitimate descriptions of physical mechanisms, capable of modelling interesting aspects of those mechanisms. The aim of this formal model is to combine a *qualitative* description of the hierarchical and causal organization of a multi-field and multi-level cancer mechanism with *quantitative* information about the strengths of the causal and hierarchical connections. In this section we examine the philosophical progress on the question of what a physical mechanism *is*, and describe how this validates our thesis.

The current dominant philosophical characterization of a mechanism is: 'Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions' (Machamer et al., 2000, p. 3). For alternative accounts see: (Glennan, 2002, p. S344); (Bechtel and Abrahamsen, 2005, p.423).

When considering the mechanisms of cancer, the *entities* are the objects involved, such as people, tumours, cells, DNA molecules, and so on. Activities are just the things these objects do, such as survive, grow, replicate, damage and be damaged, repair and be repaired, and so on. So the mechanisms of cancer can be described in terms of entities and activities: people surviving or not, tumours growing or shrinking, and DNA being damaged and being repaired.

Organization is a crucial feature of such mechanisms: 'The organization of these entities and activities determines the ways in which they produce the phenomenon' (Machamer et al., 2000, p. 3). See also (Bechtel and Abrahamsen, 2005, p.435) and (Darden, 2002, p. S355).

This is again true of cancer, with the complex organization of the human body, cells in the tumour, and all the multiple cellular mechanisms, vital to the pro-

duction of cancer.

It is now recognised that such mechanisms are usually *hierarchical*. As (Machamer et al., 2000, p. 13) write: 'Mechanisms occur in nested hierarchies and the descriptions of mechanisms in neurobiology and molecular biology are frequently multilevel. . . . lower level entities, properties, and activities are components in mechanisms that produce higher level phenomena'. That is, mechanisms involve internal structure, often several levels of internal structure.

The mechanisms of cancer involve just such hierarchical levels. There is the level of the person, their socioeconomic background, and their lifestyle choices like diet, exercise and smoking. Then there is the level of the tumour, particularly its growth rates, containment, blood supply and so on. Then the cells in the tumour have particular properties. Within the cell itself, we are increasingly able to distinguish between the levels of gene expression, RNA and proteins. All these levels affect the process of the cancer, and the patient's prognosis.

It is also recognised that explaining some phenomenon requires finding the mechanism responsible for that phenomenon using empirical work from multiple scientific fields. Craver discusses this with regard to neuroscience: 'The central idea is that neuroscience is unified not by the reduction of all phenomena to a fundamental level, but rather by using results from different fields to constrain a multilevel mechanistic explanation'—see (Craver, 2007, p. 231). See also (Russo, 2008) on social mechanisms, and (McKay-Illari and Williamson, 2009) on explanation for the spread of HIV. We can call these *multi-field* mechanisms as they are modelled on the basis of different kinds of data, for instance molecular, genetic, chemical, and environmental.

Treating cancer involves integrating information from multiple fields, each studying one of the hierarchical levels we have described. Socioeconomic, clinical, genomic, transcriptomic and proteomic data are all known to be relevant to therapy choice and prognosis.

Recursive Bayesian nets are legitimate *descriptions* of physical mechanisms, provided that intra- and inter-level relations are compatible with available theoretical knowledge. If so, RBN intra-level (causal) relations among peer variables stand for mechanisms and RBN inter-level relations stand for *decompositions* of network variables into constituent sub-mechanisms. Accordingly, simple and network variables can model entities (or sets of these entities) in their various states; and RBN causal relations can model interactions and influences among these entities, i.e., activities.

Current philosophical work on biological mechanisms tends to cover purely qualitative aspects of mechanisms ((Russo, 2008) discusses, however, quantitative modelling of social mechanisms). Our work allows the possibility of adding a quantitative dimension: the probabilities within an RBN quantify the strengths of the causal connections and lead to a joint probability distribution over all the variables in the network. Such a quantitative description of the mechanism is vital for the task of prediction, which requires determining the outcomes that are most probable given available evidence. Since causal information is required for control and mechanistic information for explanation, the RBN formalism offers the prospect of multiple uses—prediction, explanation and control—as well as the capacity to integrated different kinds of evidence and evidence from different fields.

4 CANCER APPLICATION: VARIABLES AND DATA SOURCES

In the last decade the human genome project and technological breakthroughs such as those in microarray technology and mass spectrometry-based proteomics have had a big impact on cancer research. They have led to a vast increase in available data, from genomics, transcriptomics, proteomics and metabolomics. Traditional diagnostics is under strain, and there is increasing need for biomedical decision support tools.

Bayesian nets are an obvious choice for such tools, but the prospect of being able to include hierarchy in such Bayesian nets is exciting for cancer research. Hierarchy is important to cancer research since it is still unclear how the DNA, RNA, protein and metabolic levels interact to produce cancer and affect prognosis.

Our formalism can be applied to build a model using TCGA data (clinical patient data) and NCI-60 (cell line data) integrating the following variables:

Top level: Clinical The top level includes the following recursive variables. The first recursive variable kind of tumour. (Since most studies are focused on single tumour types, only information on sub-types will be available in a single study.) Each different subtype of tumour is a value of this variable, each value corresponds to a lower-level Bayesian network of dependencies between gene expression levels. If data is available, a second important recursive variable will be metastasis. Metastasis present/absent will corre-

spond to lower-level nets of gene expression levels. Finally, it is proposed that the model will also include standard simple variables at the clinical level such as therapy, age, and survival in months.

Mid-level: DNA Damage response A first recursive variable will DNA damage response: success and failure of the DNA damage response mechanism will each be modelled by different lower-level DNA damage response nets, which consider the expression levels of various DNA damage response genes as a surrogate for DNA damage. A second recursive variable will be type of DNA damage, with values for single-strand damage, double-strand damage, and damage from alkylating agents, each corresponding to a lower-level gene expression net. There will probably be no direct measure of this, so a surrogate will have to be used. There will also be simple variables for therapy choice, assembly of repair agents, repair success and apoptosis. Apoptosis surrogates could take the form of expression levels of the Caspase-3 and Caspase-9.

Low-level Gene expression data is now readily available, and methylation status of the genes has been collected in TCGA.

Quantitatively representing both causal and hierarchical structure in a single model in this way allows extra methods of manipulating vast amounts of as-yet poorly-understood data.

5 CONCLUSION

We have introduced the recursive Bayesian network formalism, extending it from the modelling of nested causal relationships to the modelling of mechanisms. We have discussed exactly how such networks can be used to model mechanisms, thus providing an integrated modelling formalism for explanation, prediction and control. Further formal work is needed on how to perform inference in the net.

We have discussed how this formalism can be applied to modelling cancer mechanisms, where hierarchy is ubiquitous, and vast amounts of data are increasingly available. This model will be built, tested and validated in the ensuing programme of research.

We have also shown how this kind of model can add to the philosophical literature on mechanisms by integrating a quantitative description of the interaction between variables with the philosophically more familiar structural description of hierarchical relations between activities and entities.

There is promise for future theoretical work. By treating hierarchical structure in a formally equivalent way to causal structure, this formalism might allow us to extend known methods for extracting unknown causal structure from data to extracting unknown hierarchical structure. This is an exciting possibility for studying cancer, where both causal and hierarchical structure are still to be discovered in the areas opened up by new technology in the last decade.

ACKNOWLEDGEMENTS

We are grateful to the Leverhulme Trust and the British Academy for supporting this research. We are also grateful to Amos Folarin, May Yong and Sylvia Nagl of the UCL Cancer Institute for valuable discussions.

REFERENCES

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

Craver, C. F. (2007). *Explaining the brain*. Oxford University Press, Oxford.

Darden, L. (2002). Strategies for discovering mechanisms: Schema instantiation, modular subassembly, forward/backward chaining. *Philosophy of Science*, 69:S354S365.

Glennan, S. (2002). Rethinking mechanistic explanation. *Philosophy of Science. Supplement: Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association. Part II: Symposia Papers (Sep., 2002)*, 69(3):S342–S353.

Gyftodimos, E. and Flach, P. (2002). Hierarchical Bayesian networks: a probabilistic reasoning model for structured domains. In de Jong, E. and Oates, T., editors, *Proceedings of the ICML-2002 Workshop on Development of Representations*, pages 23–30. University of New South Wales.

Jaeger, M. (2001). Complex probabilistic modeling with recursive relational Bayesian networks. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):179–220.

Koller, D. and Pfeffer, A. (1997). Object-oriented Bayesian networks. In *Proceedings of the 13th Annual Conference on Uncertainty in Artificial Intelligence*, pages 302–313.

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

McKay-Illari, P. and Williamson, J. (2009). Function and organization: comparing the mechanisms of protein synthesis and natural selection. Under review.

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.

Peña, J. M., Lozano, J. A., and Larrañaga, P. (2002). Learning recursive Bayesian multinets for clustering by means of constructive induction. *Machine Learning*, 47(1):63–90.

Russo, F. (2008). *Causality and causal modelling in the social sciences. Measuring variations*. Methodos Series. Springer, New York.

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

Williamson, J. and Gabbay, D. (2005). Recursive causality in Bayesian networks and self-fibring networks. In Gillies, D., editor, *Laws and models in the sciences*, pages 173–221. King's College Publications, London. With comments pp. 223–245.