Structured probabilistic coherence and the usual counterexamples to probabilistic measures of coherence

Abstract. The notion of coherence is used in many philosophical, especially epistemological, discussions (for instance, in discussions about the truth-conduciveness of coherence), so a formal explication of this notion is desirable. Yet, such explications available on the market disagree and face a number of counterexamples. Reflecting on common phenomena that underlie these counterexamples leads us to the formulation of a new measure of coherence. It diverges from the known candidates in three important respects: (1) it is not a function of a probabilistic measure and a set of propositions alone, because it is also sensitive to the structure of agent's beliefs, (2) unlike in the case of quite a few coherence measures, it is sensitive to the weakest links in the narration, and (3) it is not obtained by simply averaging confirmation levels between all possible combinations of elements. We apply our measure to the existing counterexamples and compare its performance to the performance of the other measures. It does a better job.

1 Introduction & motivations

The notion of coherence is often used in many philosophical, especially epistemological, discussions (for instance, in discussions about the truth-conduciveness of coherence). When we talk about the coherence of a set of propositions or about the coherence of a story, we seem to refer to how well their individual pieces fit together. How are we to understand and apply this notion systematically, though? In particular, we will be interested in probabilistic explications of this notion, as Bayesian epistemology strives to be a general epistemological project and as such it should be able to accommodate coherence-oriented considerations.

There is also a more practical reason to develop a better understanding of the notion: a plausible measure of coherence could be used to better evaluate the quality of some stories or narrations. For example in the legal context we would like to be able to assess the quality of a testimony in the court of law (Allen, 2010; Pennington & Hastie, 1991; Spottswood, 2013; Vlek, 2016).

Multiple probabilistic explications of coherence have been proposed (Douven & Meijs, 2007; Fitelson, 2003; Glass, 2002; Meijs & Douven, 2007; Olsson, 2001; Roche, 2013; Shogenji, 1999). However, clear general principles to choose between them are hard to come by. One paper where some such principles have been formulated is (Schippers, 2014), where a list of seemingly plausible adequacy conditions for a coherence measure is proposed and shown to be inconsistent to argue for pluralism about the notion of coherence. However, some of those requirements are rather strong and controversial.[1]

The general point here is not that the approach taken in (Schippers, 2014) is flawed, but rather that the task of formulating general principles for coherence is a challenge, and that no clear list of such uncontroversial desiderata is on the horizon.

[1] Let us illustrate this. The (Dependence) condition formulated there requires that the coherence score of a set of propositions is above (below) the neutral score if for all pairs of non-empty subsets the posterior of an element of a pair conditional on the other element is higher than (lower than) the prior of the former. This makes some of the features of the coherence measure dependent on the priors, and whether it should be so is not obvious. On the other hand, (Agreement) is formulated in terms of conditional probabilities between such pairs. If on a given measure P all conditional probabilities (between pairs already mentioned) are higher than on P', the coherence of a set given P should be higher than given P'. The (Equivalence) requirement is that any finite set of logically equivalent propositions should be maximally coherent. This is suspicious, as the set $\{0 = 1, 2 = 5\}$ is a set of equivalent propositions (with sufficiently strong notion of logical equivalence in the background), but we would intuitively hesitate to say it's maximally coherent.

One approach to obtaining some clarity on which abstract conditions are plausible is looking at various thought experiments in which our intuitions about what the coherence scores should be (at least comparatively) are more robust than direct assessment of general requirements. In fact, looking at examples is what the main stream of literature on probabilistic coherence focused on, and each probabilistic measure of coherence faces a selection of seemingly intuitive counterexamples.

We decided to work with this methodology. We first gathered key examples that occur in the literature, developed **R** scripts calculating all coherence scores for these scenarios, expanding on the exploration already developed in Koscholke (2016).[2] Then we reflected on the results, noticing that one weakness of the measures is that they pay little attention to the underlying structure of a given narration in the calculation of its coherence.

Inspired by this observation, we formulate our own proposal, which diverges from the known purely probabilistic measures of coherence in the following important respects. (i) We don't think of agent's credal state as captured by a set of propositions and a probabilistic measure, but rather as a narration, in which the relations of (intended direct) dependence between proposition is crucial. Our representation of a narration will be composed of a set of propositions, a probability measure and a directed acyclic graph representing these structural assumptions—that is, we will be thinking of agent's credal states in terms of Bayesian networks. (ii) Coherence is not a function of a probability measure and a set of propositions alone, because it is also sensitive to this structure. (ii) Unlike in the case of quite a few coherence measures, it is sensitive to the weakest links in the narration. (iii) It is not obtained by simply averaging confirmation levels between all possible combinations of elements.

We described this approach in a more detailed introduction to this measure [ANONYMIZED], which explains the method and some of the theoretical decisions that we have made, and show how it works using a Bayesian network developed for the well-known Sally Clark case (Fenton & Neil, 2018). The goal of the current paper is to discuss a range of philosophical counterexamples to the existing probabilistic measures of coherence and evaluate the performance of our approach using those as a benchmark, arguing that it performs much better than the existing ones.

Accordingly, in Section 2 we introduce all the coherence measures, including the key motivations for and a pseudo-code description of our measure. In Section 3 we describe the thought experiments meant as counterexamples to coherence measures, their corresponding desiderata and their status on various coherence measures, including ours. The order of the discussion of any given example is straightforward: we first describe the example and the intuitive desiderata related to it, then we explain how the situation is represented by means of a Bayesian network(s), and investigate what happens when we apply all coherence measures. We end with Section 4 in which we compare all of the results and draw some general conclusions.

2 Probabilistic coherence measures and structured coherence

Quite a few different measures of coherence have been proposed in the literature. Two early proposals are:

- Shogenji's **deviation from independence** (Shogenji, 1999), is defined as the ratio between the probability of the conjunction of all claims, and the probability that the conjunction would get if all its conjuncts were probabilistically independent (scaling from 0 to ∞ with neutral point 1):

$$\mathscr{C}_S(S) = \frac{P(\bigwedge S)}{\prod_{i=1}^{n} P(S_i)} \qquad \text{(Shogenji)}$$

where $S = \{S_1, \ldots, S_n\}$ is a set of propositions whose coherence is to be evaluated. This measure was later generalized by Meijs & Douven (2007). According to this approach, (Shogenji) is applied not only to the whole set of propositions, but to each non-empty non-singleton subset of the set, and the final value is defined as the average of all sub-values thus obtained.

- **Relative overlap** (Glass, 2002; Olsson, 2001), is defined as the ratio between the intersection of all propositions and their union (scaling from -1 to 1 with no clear neutral point):

[2] As we represented the scenarios with Bayesian networks in our calculations (the choice of this form of representation is useful for computational reasons, but has no impact on the outcomes, as these outcomes are functions of probability measures), the work has been made possible by all those who contributed to the development of **R** language, and Marco Scutari, the author of **bnlearn** package, who was kind enough to extend his package with additional features upon our requests (Scutari & Denis, 2015).

$$\mathscr{C}_O(S) = \frac{P(\bigwedge S)}{P(\bigvee S)} \qquad \text{(Olsson)}$$

It has also been generalized in a way analogous to the one used in the generalization of the Shogenji's measure (Meijs & Douven, 2007).

Both of these approaches are susceptible to various objections and counterexamples (Akiba, 2000; Bovens & Hartmann, 2004; Crupi, Tentori, & Gonzalez, 2007; Koscholke, 2016; Merricks, 1995; Schippers & Koscholke, 2019; Shogenji, 1999, 2001, 2006; Siebel, 2004, 2006). To overcome them, more recent works proposed **average mutual support** measures, starting with (Fitelson, 2003). The general recipe for such measures is as follows.

- Given that S is a set whose coherence is to be measured, let P indicate the set of all ordered pairs of non-empty, disjoint subsets of S.
- First, define a confirmation function (of a hypothesis H by evidence E): $\mathsf{conf}(H,E)$.
- For each pair $\langle X,Y\rangle \in P$, calculate $\mathsf{conf}(\bigwedge X, \bigwedge Y)$, where $\bigwedge X$ is the conjunction of all the elements of X (and $\bigwedge Y$ is to be understood analogously).
- Take the mean of all the results:

$$\mathscr{C}(S) = \mathsf{mean}\left(\left\{\mathsf{conf}(\bigwedge X_i, \bigwedge Y_i) | \langle X_i, Y_i\rangle \in P\right\}\right).$$

Different measures of coherence result from different choices of a confirmation measure. Here are the key candidates present in the literature:

- Fitelson (2003) uses the following confirmation function (the resulting coherence measure ranges from -1 to 1 with neutral point at 0):

$$F(H,E) = \begin{cases} 1 & E \models H, E \not\models \bot \\ -1 & E \models \neg H \\ \frac{P(E|H)-P(E|\neg H)}{P(E|H)+P(E|\neg H)} & \text{o/w} \end{cases}$$

$$\mathscr{C}_F(S) = \mathsf{mean}\left(\left\{F(\bigwedge X_i, \bigwedge Y_i) | \langle X_i, Y_i\rangle \in P\right\}\right) \qquad \text{(Fitelson)}$$

- Douven & Meijs (2007) use the difference confirmation measure (with coherence ranging from -1 to 1 with neutral point at 0):

$$D(H,E) = P(H|E) - P(H)$$

$$\mathscr{C}_{DM}(S) = \mathsf{mean}\left(\left\{D(\bigwedge X_i, \bigwedge Y_i) | \langle X_i, Y_i\rangle \in P\right\}\right) \qquad \text{(DM)}$$

- Roche (2013) uses the absolute confirmation measure (the resulting coherence measure ranges from 0 to 1 with neutral point at 0.5):

$$A(H,E) = \begin{cases} 1 & E \models H, E \not\models \bot \\ 0 & E \models \neg H \\ P(H|E) & \text{o/w} \end{cases}$$

$$\mathscr{C}_R(S) = \mathsf{mean}\left(\left\{A(\bigwedge X_i, \bigwedge Y_i) | \langle X_i, Y_i\rangle \in P\right\}\right) \qquad \text{(Roche)}$$

Mind your head: different measures use different scales and have different neutral points (values taken for any set of probabilistically independent propositions; not all measures have neutral points). This is worth keeping in mind when it comes to various desiderata that we will discuss.

As we already mentioned in the introduction, formulating abstract formal requirements for a coherence measure and investigating whether a given coherence measure satisfies them has not resulted in an agreement. For this reason, we follow another path, which has dominated the literature on the topic. We look at how the measures behave in test scenarios. Many putative scenarios were put forward as counterexamples. They usually have the form of a few propositions formulated in natural language, such that intuitive judgments of coherence involved and the formal coherence calculations seem to diverge (Akiba, 2000; Bovens & Hartmann, 2004; Koscholke, 2016; Meijs & Douven, 2007, 2007; Merricks, 1995; Schippers & Koscholke, 2019, 2019; Shogenji, 1999, 2001, 2006; Siebel, 2004, 2006). The probabilistic measures we introduced above do not seem to handle these examples very well (read on for details).

Inspired by these failures, in [REFERENCE SUPRESSED FOR ANONYMITY] we proposed to take a different perspective. Putting the earliest measures aside (they were problematic for reasons

discussed in the literature we already referred to), we noticed that the problems with the average mutual support measures stem from the fact that the coherence score is an average confirmation score for all possible combinations of the parts of a narration. Therefore we proposed to take a more fine-grained account. First, we represented an agent's credal state by means of a Bayesian network, which comprises not only a probabilistic measure but additional structural information. Then we used this structural information in our definition of coherence, so that only those directions of support are considered which in fact are indicated by the structure of the agent's belief state.

While we refer the reader to a more extensive treatment in [REFERENCE SUPRESSED FOR ANONYMITY], we now briefly discuss the main idea behind it. A Bayesian network (BN) is a type of probabilistic graphical model that represents a set of variables and their independence relationships with a directed acyclic graph (DAG). It also involves parameters that encode probabilistic information about the variables and relations between them. Formally, a BN is defined by a pair $\langle G, \theta \rangle$, where G is a DAG whose nodes represent random variables $X_1, X_2, \ldots, X_n$, and whose edges identify direct dependencies between these variables. θ is a set of parameters that defines the factorization of a probability distribution, which includes the probability of each X_i conditional on the set of its parents (pa_i) in G: $\theta_{X_i|pa_i} = P(X_i|pa_i)$. $\theta_{X_i|pa_i}$ is called the X_i's conditional probability table (CPT).[3]

For instance, consider a DAG underlying the BN developed by (Fenton & Neil, 2018}) to illustrate a point about the notorious Sally Clark case (Figure 1).[4] The arrows depict relationships of (usually causal) influence between variables. Amurder and Bmurder are binary nodes corresponding to whether Sally Clark's sons, call them A and B, were murdered. These influence whether signs of disease (Adisease and Bdisease) and bruising (Abruising and B.bruising) were present. Also, since son A died first, whether A was murdered casts some light on the probability of son B being murdered. The notion of direct dependence, of course, is model-relative. Once we restrict attention to a set of propositions, whether dependence is direct or not depends on whether we think the influence is mediated by other variables considered in the model.

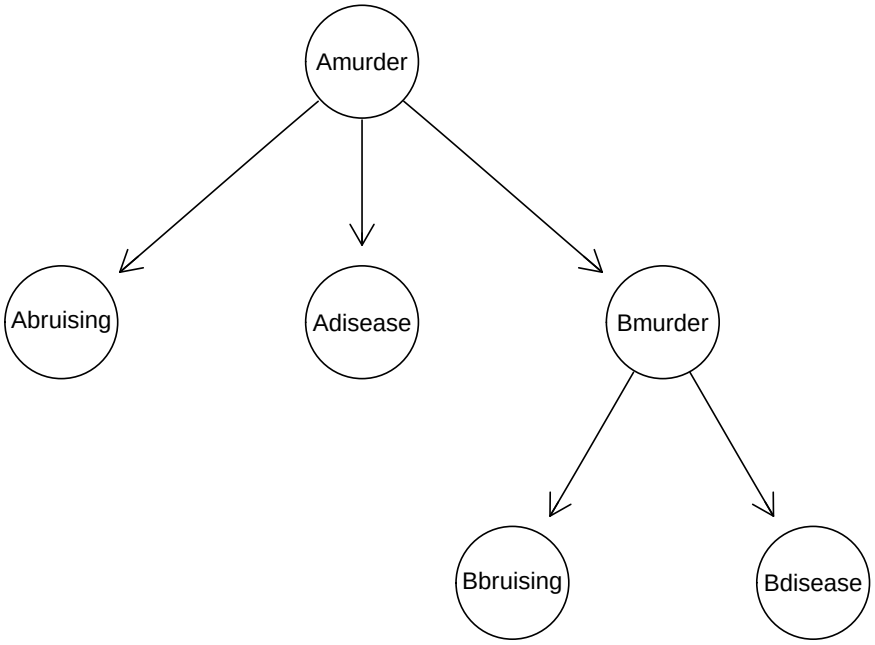


Figure 1: The directed acyclic graphs for the Sally Clark BN.

[3]We are not dealing with continuous random variables in this paper.

[4]R. v. Clark (EWCA Crim 54, 2000) is a classic example of how the lack of probabilistic independence between events can be easily overlooked. Sally Clark's first son died in 1996 soon after birth, and her second son died in similar circumstances a few years later in 1998. At trial, the paediatrician Roy Meadow testified that the probability that a child from such a family would die of Sudden Infant Death Syndrome (SIDS) was 1 in 8,543. Meadow calculated that therefore the probability of both children dying of SIDS was approximately 1 in 73 million. Sally Clark was convicted of murdering her infant sons (the conviction was ultimately reversed on appeal). The calculation illegitimately assumes independence, as the environmental or genetic factors may predispose a family to SIDS. The winning appeal was based on new evidence: signs of a potentially lethal disease—contrary to what was assumed in the original case—were found in one of the bodies.

Once the state of a node has been found out, it becomes an *evidence node*. If, for instance, the evidence includes information about the the sings of bruising and signs of disease in the children, nodes Abruising, Bbruising, Adisease and Bdisease become evidence nodes. A BN can also be put forward jointly with binary claims made by an agent. For instance, the prosecution might not only present the BN, but also claim that both children in fact have been murdered. That is, an agent might take a definite stance about a selection of nodes (here, the Amurder and Bmurder)—in such a case, these become the *narration nodes*, and the states that an agent claims they have are their corresponding *narration states*.[5]

Each parented node in the BN receives its expected confirmation score (ECS). It is calculated by looking at all combinations of its states and states of its parents not excluded by agents' narration.[6] For each of these combinations, the confirmation score between the parents' states and the child state is calculated (in the pseudo-code, we use confirmation measure Z, in further calculations we also use measures LR and L for comparison).[7] Then, a weighted average of these scores is obtained. The weights are the normalized probabilities of the combinations of parents' states obtained in the BN (updated with the evidence, if it has been specified). If both the child and the parent nodes belong to the narration, there is only one possible combination so after normalization it gets weight 1. The final coherence score is either the mean of the ecs scores for all the child nodes, if all of them are positive, or it is a weighted average of their mean and their minimum, $(1-|\mathsf{min}(\mathsf{ecs})|)\times\mathsf{mean}(\mathsf{ecs})+|\mathsf{min}(\mathsf{ecs})|\times\mathsf{min}(\mathsf{ecs})$, otherwise.[8] The appendix contains a pseudo-code for structured coherence.[9]
Having introduced the coherence measures at play, let us now move to the key counterexamples discussed in the literature.

3 Challenges and their treatment

We will now go through a list of key counterexamples proposed in the literature, each time explaining the relevant desiderata. We represent those scenarios as Bayesian networks. Then we calculate the coherence scores for those scenarios using all the measures we have introduced. Finally, we test whether the desiderata are satisfied.

Here are the counterexamples put forward against various coherence measures in the literature. We ignored only a few where both we didn't share the authors' intuitions and the examples were not picked up in further discussion in the literature.[10]

3.1 Penguins

The scenario. This is a challenge to the Olsson-Glass measure discussed in (Bovens & Hartmann, 2004, p. 50) and (Meijs & Douven, 2007). It consists of the propositions (we'll call them *nodes*, as these will be used later on in Bayesian networks) displayed in Table 1.

[5] We assume that the agent does not assign probability 0 to the narration they propose.

[6] Conceptually, it is possible to not restrict ourselves this way and look at all combinations of states of the nodes which are assigned non-null probabilities. But then, effectively, the binary content of the narration would have no impact on the coherence score, and we would obtain a coherence measure for the purely probabilistic part of agents' convictions. While this might be a worthy enterprise, we do not purse this idea in this paper. Calculations of a measure thus modified can be achieved by fairly straighforward modification of our code.

[7] Z is 0 if $\mathsf{posterior}=\mathsf{prior}$, $\frac{\mathsf{posterior}-\mathsf{prior}}{1-\mathsf{prior}}$ if $\mathsf{posterior}>\mathsf{prior}$, and $\frac{\mathsf{posterior}-\mathsf{prior}}{1-\mathsf{prior}}$ in the remaining case. Call the posterior of the evidence given a hypothesis likelihood and the probability of the evidence conditional on the negation of the hypothesis nlikelihood. Then, likelihood ratio (LR) is $\frac{\mathsf{likelihood}}{\mathsf{nlikelihood}}$ and L is $\frac{\mathsf{likelihood}-\mathsf{nlikelihood}}{\mathsf{likelihood}+\mathsf{nlikelihood}}$

[8] We have developed **R** code calculating this and other measures to handle calculations that will be discussed further on, the code with documentation is available at ANONYMIZED.

[9] For other confirmation measures the code needs to be modified, but the required modification is straightforward).

[10] One such an example, involves Sarah and her pregnancy (Shogenji, 2006), but it focused more on truth-conduciveness of coherence, which is beyond the scope of our paper. We also do not discuss a few other examples involving fossils and voltage (Shogenji, 2001; Siebel, 2006). In some respects, they were quite similar to the dice and depth problems that we do discuss, and some of their variants simply did not inspire our agreement. For instance, Siebel thinks that for voltage levels $\{V=1,V=2\}$ is more coherent than $\{V=1,V=50\}$, while we think that both sets are maximally incoherent (there might be some claims in the vicinity that are not incoherent, say, focusing on results of separate measurements, but an example along these lines has not been properly formulated in the literature).

node	content
B	Tweety is a bird.
G	Tweety is a grounded animal.
P	Tweety is a penguin.

Table 1: Propositions in the Penguins scenario

Desiderata. Meijs & Douven (2007) claim that the set {B,G}, which doesn't contain the information about Tweety being a penguin, should be less coherent than the one that does contain this information: {B,G,P}.

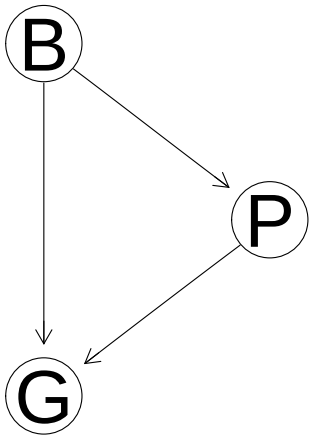
(BG<BGP) {B,G} should be less coherent than {B,G,P}.

Another intuition about this scenario (Schippers & Koscholke, 2019) is that when you consider a set which says that Tweety is both a bird and a penguin: {B,P}, adding proposition about not flying (G) shouldn't increase the coherence of the set as much as moving from {B,G} to {B,G, P}. It's a well-known fact that penguins don't fly and by adding G explicitly to the set, one wouldn't gain as much information. However, as G is not a logical consequence of P, it can be argued that {B,P} and {B,P,G} represent different information sets, and so some difference in their coherence is to be expected.

(BG $\ll$ BP$\leq$BGP) {B,P} should be notably above {B,G}, and less than {B,P,G}.

Formally, we'll require that the absolute difference between BG and BP be greater than .1 (the exact placement of the threshold doesn't make a huge difference, unless it's at an unintuitive value below .01) and that {B,G} $\leq$ {B,P,G}.

Bayesian network. We used the distribution used in the original formulation (Meijs & Douven, 2007) to build a BN corresponding to the narrations at play (Fig. 2).[11]



B	Pr
1	0.5
0	0.5

P	B	
	1	0
1	0.02	0
0	0.98	1

	B	P	
G			Pr
1	1	1	1.00
0	1	1	0.00
1	0	1	0.00
0	0	1	1.00
1	1	0	0.00
0	1	0	1.00
1	0	0	0.98
0	0	0	0.02

Figure 2: Bayesian network for the Penguins problem.

Results. Now, let's calculate the coherence scores (Table 2) and see if the desiderata are satisfied (Table 4). The measures are: Olsson-Glass (OG), generalized Olsson-Glass (OGgen), Shogenji (Sh), generalized Shogenji (ShGen), Douven-Meijs (DM), Roche (R), Fitelson (Fi), Structured with Z (SZ), LR (SLR), and L (SL) used as a confirmation measure.[12]

[11] Not without concerns. There are around 18 000 species of birds, and around 60 of them are flightless. We couldn't find information about counts, but it seems the probability of being a penguin if one is grounded is overestimated by philosophers. Also, there are many things that are not grounded but are not birds, mostly insects, and there's plenty of them. We did spend some time coming up with plausible ranges of probabilities to correct for such factors, and none of them actually makes a difference to the main point. So, for the sake of simplicity, we leave the original unrealistic distribution in our discussion.

[12] One phenomenon woth noticing is that the structured measure in its variant that employs likelihood ratio gives Infinity in one case. The reason why this happens is as follows. One of the child nodes is G. We need to calculate the expected support its parents' states provide for G. In the case the scenario actually specifies that all nodes have value 1, this boils down to calculating the likelihood ratio for only one case. We treat $(\mathsf{B}=1 \wedge \mathsf{P}=1)=\mathsf{E}$ as the evidence. $\mathsf{P}(E|\mathsf{G}=1)=.02$, that is, if

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Penguins: BGP 111	0.01	0.015	4.00	2.01	0.255	0.51	0.453	0.505	Inf	0.669
Penguins: BG 11	0.01	0.010	0.04	0.04	-0.480	0.02	-0.960	-0.960	0.02	-0.960
Penguins: BP 11	0.02	0.020	2.00	2.00	0.255	0.51	0.669	0.010	2.02	0.338

Table 2: Coherence scores for the Penguins scenario (rounded). Note how LR might result in Inf if a conditional probability of 1 in the calculations is involved.

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Penguins: BG<BGP	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Penguins: BG<< BP< BGP	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Table 3: Desiderata satisfaction for the Penguins scenario.

3.2 Dice

The scenario. This scenario was offered by Schippers & Koscholke (2019). You're either tossing a regular die, or a dodecahedron, X is the result (there is nothing particular about this choice of dice; *mutatis mutandis* this should hold for other possible pairs of dice as well). Consider the coherence of the set whose first element says that the result of the toss is a two, and whose second element says that the result of the toss is either a two or a four. $D = \{X = 2, (X = 2 \vee X = 4)\}$.

Desiderata. According to Schippers & Koscholke (2019), in this scenario posterior conditional probabilities are fixed: getting 2 or 4 logically follows from getting 2 ($P(X = 2 \vee X = 4|X = 2) = 1$), and you always have 50% chance to get 2 given that the outcome was 2 or 4 ($P(X = 2|X = 2 \vee X = 4) = 0.5$). Therefore, they claim, the coherence of the set D should not change no matter which die you use.

(D=const) the coherence of D should not change.

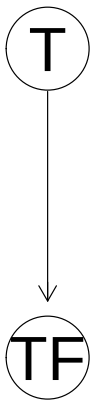
Bayesian networks. When one is about to represent the scenario as a BN the situation is not straightforward. On one hand, the scenario description includes probabilistic information about what die we are dealing with, but on the other, the probabilities the authors explicitly bring up (1 and .5) do not capture this information. The simplest approach is to build a DAG containing two nodes.

node	content
TF	The result is a two or a four.
T	The result is a two.

Table 4: Propositions in the dice scenario

and since TF logically follows from T, it's natural to draw an arrow from the latter to the former. Then, the BNs corresponding to different types of dice will share the DAG, and will differ in their CPTs. These simple BNs do not model the information about which die is tossed explicitly (as a separate node belonging to the narration, but rather the choice of the die is treated as a background information that impacts the states and probabilities of T and TF).

Tweety is grounded, the probability that it's a bird and a penguin is quite low, but $\mathsf{P}(E|\neg\mathsf{G}) = 0$, because it is impossible that Tweety is a bird and a penguin if she isn't grounded. But then, the LR would require dividing .02 by 0, which the calculations take at the limit to be infinity. Such cases will come up whenever some combinations of evidence under considerations are excluded by the state of a child node, and in the expected value calculations the presence of infinity will spread, as it is enough that it's the LR for at least one combination with a non-zero probability. This is one of the reasons we do not really think likelihood ratio is the appropriate measure for coherence calculations.



T	Pr
1	1/6
0	5/6

TF	T	
	1	0
1	1	1/5
0	0	4/5

T	Pr
1	1/12
0	11/12

TF	T	
	1	0
1	1	1/11
0	0	10/11

Figure 3: Two-node BNs for the dice problem, with CPTs for a regular die (center) and for a dodecahedron (right)

Results. If we do so, the results are as expected for structured coherence—not so much for Shogenji, generalized Shogenji, Douven-Meijs and Fitelson (Tables 5 and **??**). As the child node logically follows from the parent node, the structured coherence score is 1 if we use Z or L as the confirmation measure, and ∞ if we use the likelihood ratio.

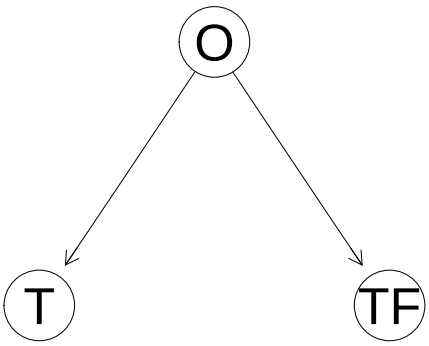
	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Dice: TTF 11	0.5	0.5	3	3	0.500	0.75	0.833	1	Inf	1
Dice: TTF 111	0.5	0.5	6	6	0.625	0.75	0.917	1	Inf	1

Table 5: Coherence scores for a regular die and a dodecahedron in the dice problem with two nodes (rounded).

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Dodecahedron: Regular = Dodecahedron	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	NA	TRUE

Table 6: Desideratum satisfaction for two-node BNs in the dice problem.

Bayesian networks (again). However, if we prefer to explicitly include the probabilistic information about all the potential outcomes as part of agents' overall narration considered in our coherence evaluation, we can use a third node to represent the outcome of the toss. We do so by adding a node O with as many equally probable states as sides of the die in question. Then, again, the DAG will be shared by two BNs, which will differ in their CPTs depending on which die we're dealing with.



O	Pr
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

O	Pr
1	1/12
2	1/12
3	1/12
4	1/12
5	1/12
6	1/12
7	1/12
8	1/12
9	1/12
10	1/12
11	1/12
12	1/12

T	O					
	1	2	3	4	5	6
1	0	1	0	0	0	0
0	1	0	1	1	1	1

T	O											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1	0	0	0	0	0	0	0	0	0	0
0	1	0	1	1	1	1	1	1	1	1	1	1

TF	O					
	1	2	3	4	5	6
1	0	1	0	1	0	0
0	1	0	1	0	1	1

TF	O											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0	1	0	1	0	0	0	0	0	0	0	0
0	1	0	1	0	1	1	1	1	1	1	1	1

Figure 4: Three-node BNs for the dice problem, with CPTs for a regular die and for a dodecahedron.

Results (again). In such a setup, the results are as in Table 7, and the status of the desideratum is summarized in Table 8.

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Dice (three nodes): TTF 11	0.5	0.5	3	3	0.500	0.75	0.833	-0.611	Inf	-0.611
Dice (three nodes): TTF 111	0.5	0.5	6	6	0.625	0.75	0.917	-0.819	Inf	-0.819

Table 7: Coherence scores for a regular die and a dodecahedron in the dice problem with three nodes (rounded).

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Dodecahedron: Regular = Dodecahedron	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	FALSE	NA	FALSE

Table 8: Desideratum satisfaction for three-node BNs in the dice problem.

This might seem problematic, but really should not. After all, now the narration represented by the three-node version is not merely 'the result is a two (and a two or a four)', but rather 'I will toss this n-sided fair die and the result will be a two (and a two or a four).' Both BNs models are based on the same agent's probabilistic convictions, but the narrations captured by the two representations are different.

If O is included and all possible results of a toss are equally probable, the outcome being two is actually not that likely, so there is no surprise the coherence score is negative (the more negative the larger the n) if the probabilistic information related to O is part of the narration.[13]

We could also explore the option of the narration being 'the die is fair, it was tossed and the result was a two (and a two or a four).' The difference here is that now the agent no longer thinks that all the states of O are all equally likely, O is updated on the evidence (which is T). When we proceed this way, the structured coherence scores are 1s.

[13] Notice also how the measure that uses LR fails to pick up on this difference, as at least one of confirmation scores in the calculations is ∞ and this fact overwhelms the expected confirmation score calculations.

3.3 Dunnit

The scenario. Another challenge, introduced by Merricks (1995) goes as follows: Mr. Dunnit is a suspect in the murder case. Detectives first obtained the body of evidence specified in Table 9.

node	content
I	Witnesses claim to have seen Dunnit do it (incriminating testimony).
M	Dunnit had a motive for the murder.
W	A credible witness claims to have seen Dunnit two hundred miles from the scene of the crime at the time of the murder.

Table 9: Initial evidence in the Dunnit scenario.

In light of this information they try to assess whether Dunnit is responsible for the crime (Table 10).

node	content
G	Dunnit is guilty.

Table 10: The guilt statement in the Dunnit scenario.

Now, suppose the detectives learn Dunnit has a twin brother (Table 11).

node	content
Tw	Dunnit has an identical twin which was seen by the credible witness two hundred miles from the scene of the crime during the murder.

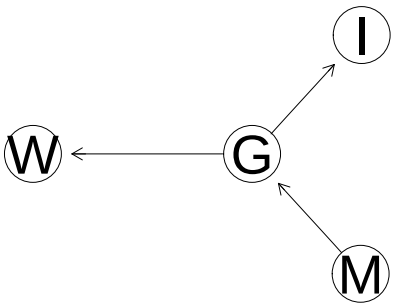
Table 11: New evidence in the Dunnit scenario.

What are our intuitions when we compare the coherence of $\{\mathsf{I},\mathsf{M},\mathsf{W},\mathsf{G}\}$ with the coherence of $\{\mathsf{I},\mathsf{M},\mathsf{W},\mathsf{G},\mathsf{Tw}\}$?

Desideratum. It seems that adding proposition about a twin should increase the coherence of the set.

(Dunnit<Twin) $\{\mathsf{I},\mathsf{M},\mathsf{W},\mathsf{G}\}$ should be less coherent than $\{\mathsf{I},\mathsf{M},\mathsf{W},\mathsf{G},\mathsf{Tw}\}$.

Bayesian networks. Here, we deal with two separate BNs. One, before the Twin node is even considered (Figure 5), and one with the Twin node (Figure 6). The CPTs for the no-twin version are in agreement with those in the ones in the Twin case. Since Merricks (1995) did not specify probabilities in the original example (and its further discussion focused on the general relationship between coherence and the addition of propositions), we came up with plausible values.



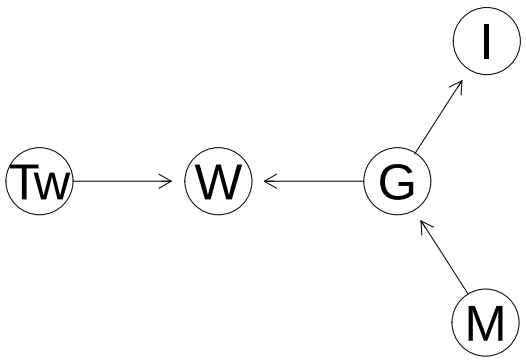
M	Pr
1	0.4
0	0.6

G	M	
	1	0
1	0.05	0.005
0	0.95	0.995

I	G	
	1	0
1	0.8	0.005
0	0.2	0.995

W	G	
	1	0
1	0.012	0.207
0	0.988	0.793

Figure 5: Twin-less BN for the Dunnit problem.



Tw	Pr
1	0.034
0	0.966

W	G	Tw	Pr
1	1	1	0.200
0	1	1	0.800
1	0	1	0.400
0	0	1	0.600
1	1	0	0.005
0	1	0	0.995
1	0	0	0.200
0	0	0	0.800

Figure 6: BN for the Dunnit problem. The key difference for the twin version lies in the construction of the CPT for W. The table gives conditional probabilities for W given various joint states of Tw and G.

Results. Coherence calculations result in Table 12 and how they fare with respect to the desideratum is displayed in Table 13.

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Dunnit: MGWI 1111	0	0.087	4.294	11.012	0.167	0.266	0.169	-0.891	56.689	-0.817
Dunnit: MTwGWI 11111	0	0.042	73.836	13.669	0.150	0.214	0.385	0.267	57.002	0.451

Table 12: Coherence scores for the Dunnit scenario (rounded).

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Dunnit: Dunnit<Twin	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE

Table 13: Desideratum satisfaction for the Dunnit scenario.

3.4 Japanese swords

The scenario. The next challenge comes from (Meijs & Douven, 2007, p. 414):

> We start by considering two situations in both of which it is assumed that a murder has been committed in a street in a big city with 10,000,000 inhabitants, 1,059 of them being Japanese, 1,059 of them owning Samurai swords, and 9 of them both being Japanese and owning Samurai swords. In situation I we assume that the murderer lives in the city and that everyone living in the city is equally likely to be the murderer. In situation II, on the other hand, we make the assumption that the victim was murdered by someone living in the street in which her body was found. In that street live 100 persons, 10 of them being Japanese, 10 owning a Samurai sword, and 9 both being Japanese and owning a Samurai sword. [...] [In situation III] we have 12 suspects who all live in the same house, and 10 of them are Japanese, 10 own a Samurai sword, and 9 are both Japanese and Samurai sword owners.

The nodes involved are as in Table 14.

node	content
J	The murderer is Japanese.
O	The murderer owns a Samurai sword.

Table 14: Nodes in the Japanese swords scenario.

Now, we look at three separate scenarios: (1) The murderer lives in the city, (2) The murderer lives

in the street popular among Japanese owners of Samurai swords, and (3) The murderer lives in the house with many Japanese owners of Samurai swords.

Desiderata. In all of the above situations the number of Japanese owners of Samurai swords remains the same. However, situations 1 and 2 differ in the relative overlap of J and O. Because J and O are more correlated in situation 2, it seems more coherent than situation 1.

(JO2>JO1) {J,O,2} should be more coherent than {J,O,1}.

However, bigger overlap doesn't have to indicate higher coherence. In situation 3 J and O confirm each other to a lesser extent than in situation 2 (compare $P(J|O) - P(J)$ and $P(O|J) - P(O)$ in both cases), and for this reason Douven and Meijs claim that situation 2 is more coherent than situation 3.

(JO2>JO3) {J,O,2} should be more coherent than {J,O,3}.

Bayesian networks. There is a common DAG for the three scenarios, but the CPTs differ (Figure 7).

J	Pr
1	0
0	1

O	J	
	1	0
1	0.008	0
0	0.992	1

(a) Scenario 1.

J	Pr
1	0.1
0	0.9

O	J	
	1	0
1	0.9	0.011
0	0.1	0.989

(b) Scenario 2.

J	Pr
1	0.833
0	0.167

O	J	
	1	0
1	0.9	0.5
0	0.1	0.5

(c) Scenario 3.

Figure 7: A common DAG and three sets of CPTs for the Japanese Swords problem.

Results. Coherence calculations are in Table 15 and the status of the desiderata involved is displayed in Table 16.

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Japanese Swords 1: JO 11	0.004	0.004	80.251	80.251	0.008	0.008	0.976	0.008	80.93	0.976
Japanese Swords 2: JO 11	0.818	0.818	9.000	9.000	0.800	0.900	0.976	0.889	81.00	0.976
Japanese Swords 3: JO 11	0.818	0.818	1.080	1.080	0.067	0.900	0.286	0.400	1.80	0.286

Table 15: Coherence scores in the Japanese swords scenarios (rounded).

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Swords: JO2>JO1	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Swords: JO2>JO3	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE

Table 16: Desiderata satisfaction in the Japanese swords scenarios.

3.5 Robbers

The scenario. A challenge put forward by Siebel (2004, p. 336) goes as follows:

> Let there be ten equiprobable suspects for a murder. All of them previously committed at least one crime, two a robbery, two pick-pocketing, and the remaining six both crimes. There is thus a substantial overlap: of the total of eight suspects who committed a robbery, six were also involved in pick-pocketing, and conversely.

The nodes involved are Table 17.

node	content
W	Real perpetrator status (states: OnlyP, OnlyR, Both).
P	The murderer is a pickpocket.
R	The murderer is a robber.

Table 17: Nodes in the Robbers scenario.

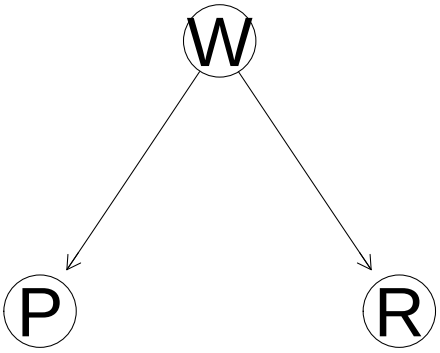
Desiderata. The first observation is that the set of propositions that corresponds to the situation in which a murderer committed both crimes should be regarded coherent. Most suspects committed both crimes, so this option is even the most probable one.

(PR>neutral) {P,R} should be regarded coherent.

According to Siebel (2004, p. 336) committing both crimes by the murderer should also be regarded more coherent than committing only one crime.

(PR>P¬R) {P,R} should be more coherent than {P,¬R} and {¬P,R}.

Bayesian networks. The robbers counterexample illustrates the point we already discussed when we talked about the Dice example: a purely probabilistic assumption may or may not be explicitly included as part of the narration. Its explicit inclusion, achieved by the addition of a node in a BN can play a role in coherence calculations. Here, such an inclusion is in line with the example as it was proposed in the literature, where the underlying prior probabilities were explicitly listed as part of the story.



W	Pr
OnlyP	0.2
OnlyR	0.2
Both	0.6

P	W		
	OnlyP	OnlyR	Both
1	1	0	1
0	0	1	0

R	W		
	OnlyP	OnlyR	Both
1	0	1	1
0	1	0	0

Figure 8: BN for the Robbers problem.

Results. Coherence calculations yield the results in Table 18, and the performance of the coherence measures with respect to the desiderata is illustrated in Table 19.

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Robbers: PR 11	0.60	0.60	0.937	0.937	-0.050	0.750	-0.143	0.60	Inf	0.60
Robbers: PR 10	0.25	0.25	1.250	1.250	0.125	0.625	0.571	-0.36	Inf	-0.36
Robbers: PR 01	0.25	0.25	1.250	1.250	0.125	0.625	0.571	-0.36	Inf	-0.36

Table 18: Coherence scores in the Robbers scenario (rounded).

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Robbers: PR>P¬R	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE
Robbers: PR>neutral	NA	NA	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE

Table 19: Desiderata satisfaction in the Robbers problem.

3.6 The Beatles

The scenario. The challenge has been offered by Shogenji (1999, p. 339) to criticize defining coherence in terms of pairwise coherence — it shows there are jointly incoherent pairwise coherent sets. The scenario consists of the claims displayed in Table 20.

node	content
D	Exactly one of the Beatles (John, Paul, George and Ringo) is dead.
J	John is alive.
P	Paul is alive.
G	George is alive.
R	Ringo is alive.

Table 20: Nodes in the Beatles scenario.

Desiderata. The set consisting of all of these propositions is logically inconsistent (even though the propositions are pairwise consistent), so it seems quite intuitive that it should be incoherent.[14]

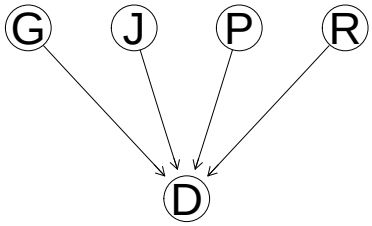
(below neutral) {D,J,P,G,R} should be incoherent.

We can make this desideratum a bit stronger by requiring that the coherence score for {D,J,P,G,R} should be minimal.

(minimal) {D,J,P,G,R} should get the lowest possible coherence value.

One of the important features of the example is that it illustrates the behavior of a coherence measure with respect to logical inconsistency. Notably, the set is logically inconsistent, and yet it will turn out that some coherence measures will assign to it non-minimal or even above-neutral coherence scores.

Bayesian network. For the sake of example, we assume the prior probability of each individual band member being dead to 0.5 (as in the above table), and the CPT for D is many-dimensional and so difficult to present concisely, but the method is straightforward: probability 1 is given to D in all combinations of the parents in which exactly one is true, and otherwise D gets conditional probability 0.



G	Pr
1	0.5
0	0.5

Figure 9: Bayesian network for the Beatles scenario.

Results. Coherence calculations give the results from Table 21, and the satisfaction of the desiderata involved can be inspected by looking at Table 22.

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Beatles: JPGRD 11111	0	0.202	0	1.423	0.025	0.322	-0.036	-1	0	-1

Table 21: Coherence scores in the Beatles scenario (rounded).

[14] One may argue that some coherence measures also measure the degree of incoherence, therefore logically inconsistent sets don't need to get the minimal score. We do not focus on such an understanding of coherence in this paper. If you think different inconsistent scenarios can differ in coherence—in line with (Easwaran & Fitelson, 2015)—our measure can accommodate this move by revising the calculations of coherence based on the ecs scores (for example, the penalty for the weakest link can be lowered, or dropped).

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Beatles: below neutral	NA	NA	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
Beatles: minimal	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE

Table 22: Desiderata satisfaction in the Beatles scenario.

3.7 The Witnesses

The scenario. This example comes from (Olsson, 2005, p. 391). Equally reliable witnesses try to identify a criminal. Consider the reports listed in Table 23 (we extended the original scenario by adding W5).

node	content
W1	Witness no. 1: "Steve did it"
W2	Witness no. 2: "Steve did it"
W3	Witness no. 3: "Steve, Martin or David did it"
W4	Witness no. 4: "Steve, John or James did it"
W5	Wittness no. 5: "Steve, John or Peter did it"
D	Who committed the deed (6 possible values)

Table 23: Testimonies in the Witnesses scenario.

Olsson, when he originally brought up the example, focused on the content of the testimonials arguing that {''Steve did it'','Steve did it"} and {"Steve, Martin or David did it", "Steve, John or James did it"} are logically equivalent (presumably, with the assumption that exactly one person did it), and so will be assigned the same level of coherence, which he finds unintuitive.

On our construal, the intuition that there is a difference in coherence between the two sets of testimonies essentially depends on the fact that they are testimonies. We take each proposition to have the structure ' 'Witness no. X claims that ..." instead—but then, the sets are no longer logically equivalent and a non-trivial probability measure can be used in coherence calculations.

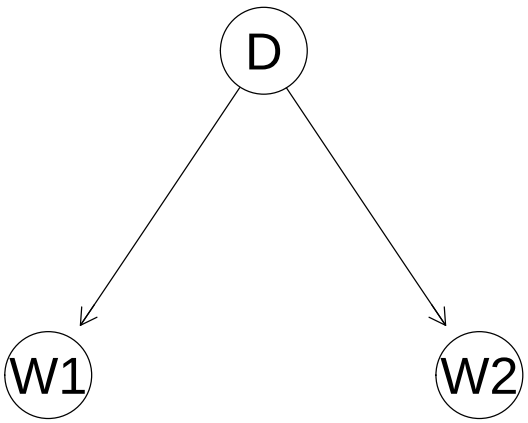
Desiderata. First, we can observe that W1 and W2 fully agree. Testimonies of W3 and W4 overlap only partially, therefore it seems that {W1,W2} is more coherent than {W3,W4}.

(W1W2>W3W4) {W1,W2} should be more coherent than {W3,W4}.

Similarly, there is a greater agreement between W4 and W5 than W3 and W4, so {W4,W5} seems more coherent than {W3,W4}.

(W4W5>W3W4) {W4,W5} should be more coherent than {W3,W4}.

Bayesian networks. The basic idea behind the CPTs we used is that for any particular witness we take the probability of them including the perpetrator in their list to be .8, and the probability of including an innocent to be .05. Of course, the example can be run with different conditional probability tables. Moreover, in this case, the fact that the witnesses provided their testimonies constitutes evidence (in contrast with the Robbers scenario, where there is no evidence as to who the perpetrator is), and so we update on it in our weights calculations. Let's first take a look at the BN for the {W1,W2} (Figure 10).



D	Pr
Steve	0.167
Martin	0.167
David	0.167
John	0.167
James	0.167
Peter	0.167

W1	D					
	Steve	Martin	David	John	James	Peter
1	0.8	0.05	0.05	0.05	0.05	0.05
0	0.2	0.95	0.95	0.95	0.95	0.95

Figure 10: BN for the W1W2 narration in the Witness problem. CPT for W2 is identical to the one for W1.

In the remaining two BNs for the problem the CPT for D remains the same, and the CPTs for the witness nodes are analogous to the one for W1. The remaining BNs have the following DAGs (Fig. 11).

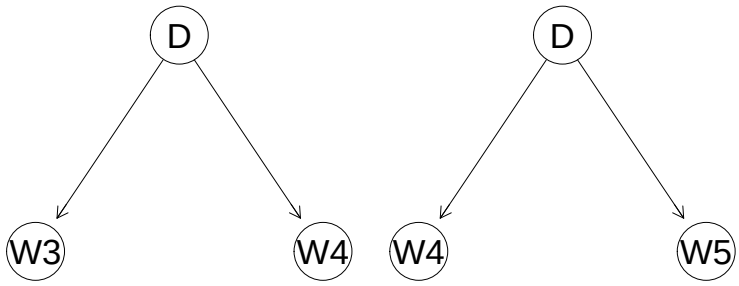


Figure 11: Two remaining DAGs for the Witness problem.

We think that what this example illustrates is that we should really carefully think about whose cognitive perspective is taken when we represent a narration using a BN, focusing on whether the BN involves nodes which are not a part of the narration whose coherence is to be evaluated. In particular, the probabilistic information about the uniform distribution of guilt probability is not a part of any of the three involved narrations, but rather a part of a third-person set-up prior to obtaining any evidence.

To evaluate the coherence of a narration, one should think counterfactually, granting the (potentially probabilistic) consequences of the narration. In our case, a judge who evaluates the coherence of witness testimonies once she has heard them, no longer thinks that the distribution of D is uniform. And this agrees with the counterfactual strategy we just described: it is a consequence of the probabilistic set-up and the content of W1 and W2 that if W1 and W2 were true, the distribution for D no longer would be uniform, and so it is unfair to judge the coherence of this scenario without updating one's assumptions about D.

Results. Note that in this case we're dealing with the perspective of someone who starts with a uniform prior for D, and subsequently considers what would happen if she obtained a given set of testimonies as evidence. Clearly, then, the distribution for D would no longer be uniform, but rather result from updating on this evidence (the posterior is presented in Table 24).

	Steve	Martin	David	John	Peter
Pr	0.981	0.004	0.004	0.004	0.004

Table 24: Propagated probabilities for D in the Witnesses scenario (rounded).

In the coherence calculations we first obtain how specific states of the narration child nodes would be confirmed by various states of D. In calculating these confirmation scores we do *not* use the updated BN (recall the old evidence problem: confirmation by evidence already included would be null). However, the weights assigned to confirmation scores in the ecs calculations should be the (normalized) probabilities obtained by updating on the evidence. If given the evidence included in the scenario Steve is the most likely perpetrator, in your coherence calculations you give the most weight to the confirmation score obtained if he indeed is the perpetrator.

Once we make a distinction between the logically equivalent sets of contents of testimonies, and non-trivially dependent sets of statements about which witness said what, the problem turns out to be not that challenging for any of the coherence measures under discussion with the prior we described. The coherence scores are displayed in Table 25 and the status of the desiderata is in Table 26.

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Witness: W1W2 11	0.451	0.451	3.551	3.551	0.446	0.621	0.771	0.729	18.501	0.871
Witness: W3W4 11	0.187	0.187	0.740	0.740	-0.110	0.315	-0.234	0.494	4.863	0.528
Witness: W4W5 11	0.365	0.365	1.260	1.260	0.110	0.535	0.218	0.602	5.236	0.637

Table 25: Coherence scores in the Witnesses scenario (rounded).

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Witness: $W_1W_2 > W_3W_4$	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Witness: $W_4W_5 > W_3W_4$	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Table 26: Desiderata satisfaction in the Witnesses scenario.

3.8 Depth

The scenario. There are eight equally likely suspects $1,\ldots,8$, and three equally reliable witnesses each trying to identify the person responsible for the crime. Compare two sets, X1 and X2, of claims about who is responsible. In contrast with Witnesses, the example ignores the fact that these are testimonies and focuses on the material disjunctive content:

$$X_1 = \{(1 \vee 2 \vee 3), (1 \vee 2 \vee 4), (1 \vee 3 \vee 4)\}$$
$$X_2 = \{(1 \vee 2 \vee 3), (1 \vee 4 \vee 5), (1 \vee 6 \vee 7)\}$$

Desiderata. In X1 witnesses' testimonies have bigger overlap, between each pair of the witnesses 2 suspects are the same, and in X2 only 1 suspect is always the same. Following Schupbach (2008), one may have an intuition that the first situation is more coherent.

(X1>X2) X_1 should be more coherent than X_2.

This desideratum presupposes we are dealing with claims made by different agents. However, on our approach the narration nodes and the BN involved are supposed to represent one agent's credal state. As long as you take the narration to be equivalent to the conjunction of the elements, an agent accepting all the elements of either X_1 or X_2 would simply accept that 1 is responsible, and we do not think it makes sense to ask about the coherence of a single proposition. Moreover, since both sets are logically equivalent, the desideratum would no longer be intuitive on the single-agent approach.

There is a way to represent the fact that different people made different testimonies within our approach by preceding the claims with "witness a said". We already employed this strategy to the Witnesses scenario, and will not do this here again, as the problem structure is essentially the same.

If we insist that the disjunctions are put forward by different agents and the agent whose perspective we are trying to model does not assume that all of them are true, we should turn to another tool. There already exist working measures of such an agreement which give the desired results. For instance, both Fleiss κ and Light's κ (Fleiss, 1971) give 0.467 for X1 and -0.067 for X2, which is in line with the desideratum.

4 Summary

Our goal was to improve on the existing probabilistic approaches to the notion of coherence. The main problem we identified had to do with taking average confirmation for all possible combinations of the elements of a given narration without paying attention to its structure. Accordingly, we developed an approach on which a narration is represented by means of a Bayesian network which captures additional structural information, and a selection of nodes on which an agent has a decisive stance. Given such a representation, the coherence of a narration is—roughly speaking—a function of the expected support of children nodes in the network, ignoring combinations of states logically excluded by the narration.

In the literature there is a tradition of criticizng various probabilistic explications of coherence from the perspective of philosophical thought experiments meant to pump our intuitions about what conditions coherence should satisfy. We followed this path, arguing that our measure copes with such counterexamples much better than the other candidates on the market. All the calculations are displayed in Table **??** and the desiderata yield Table **??**, with corresponding success rates in the last row.

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Penguins: BGP 111	0.010	0.015	4.000	2.010	0.255	0.510	0.453	0.505	Inf	0.669
Penguins: BG 11	0.010	0.010	0.040	0.040	-0.480	0.020	-0.960	-0.960	0.020	-0.960
Penguins: BP 11	0.020	0.020	2.000	2.000	0.255	0.510	0.669	0.010	2.020	0.338
Dice: TTF 11	0.500	0.500	3.000	3.000	0.500	0.750	0.833	1.000	Inf	1.000
Dice: TTF 111	0.500	0.500	6.000	6.000	0.625	0.750	0.917	1.000	Inf	1.000
Dunnit: MGWI 1111	0.000	0.087	4.294	11.012	0.167	0.266	0.169	-0.891	56.689	-0.817
Dunnit: MTwGWI 11111	0.000	0.042	73.836	13.669	0.150	0.214	0.385	0.267	57.002	0.451
Japanese Swords 1: JO 11	0.004	0.004	80.251	80.251	0.008	0.008	0.976	0.008	80.930	0.976
Japanese Swords 2: JO 11	0.818	0.818	9.000	9.000	0.800	0.900	0.976	0.889	81.000	0.976
Japanese Swords 3: JO 11	0.818	0.818	1.080	1.080	0.067	0.900	0.286	0.400	1.800	0.286
Robbers: PR 11	0.600	0.600	0.937	0.937	-0.050	0.750	-0.143	0.600	Inf	0.600
Robbers: PR 10	0.250	0.250	1.250	1.250	0.125	0.625	0.571	-0.360	Inf	-0.360
Robbers: PR 01	0.250	0.250	1.250	1.250	0.125	0.625	0.571	-0.360	Inf	-0.360
Beatles: JPGRD 11111	0.000	0.202	0.000	1.423	0.025	0.322	-0.036	-1.000	0.000	-1.000
Witness: W1W2 11	0.451	0.451	3.551	3.551	0.446	0.621	0.771	0.729	18.501	0.871
Witness: W3W4 11	0.187	0.187	0.740	0.740	-0.110	0.315	-0.234	0.494	4.863	0.528
Witness: W4W5 11	0.365	0.365	1.260	1.260	0.110	0.535	0.218	0.602	5.236	0.637

Table 27: Coherence scores in all the examples considered (rounded).

	OG	OGGen	Sh	ShGen	DM	R	Fi	SZ	SLR	SL
Penguins: BG<BGP	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Penguins: BG<< BP< BGP	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Dodecahedron: Regular = Dodecahedron	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	NA	TRUE
Dunnit: Dunnit<Twin	FALSE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
Swords: JO2>JO1	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Swords: JO2>JO3	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE
Robbers: PR>P¬R	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE
Robbers: PR>neutral	NA	NA	FALSE	FALSE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE
Beatles: below neutral	NA	NA	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	TRUE
Beatles: minimal	TRUE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE
Witness: $W_1W_2 > W_3W_4$	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Witness: $W_4W_5 > W_3W_4$	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
Success rate	50%	50%	67%	50%	50%	67%	67%	100%	91%	100%

Table 28: Desiderata satisfaction in all the examples considered, including success rates.

Appendix: structured coherence calculation pseudo-code

```
FUNCTION parents_child_possible_states(parents,child,narration)
    IF child included in narration THEN
        consequentStates <- the unique state of child as reported in the narration
    ELSE
        consequentStates <- all possible states of child
    FOR EACH parent IN parents
        IF parent included in narration THEN
            parentStates[parent] <- the unique state of parent as reported in the narration
        ELSE
            parentStates[parent] <- all possible states of parent
    parentsStates <- all combinations of parentStates
    variants <- list of all possible combinations of consequentStates and parentsStates
    RETURN variants

FUNCTION coherence_from_ecs(ecs)
    IF min(ecs) <= 0 THEN
        RETURN mean(ecs) * (min(ecs)+1) - min(ecs)min(ecs)
        #this is equivalent to (1- |min(ecs)|) * mean(ecs) + |min(ecs)| * min(ecs)
    ELSE
        RETURN mean(ecs)

FUNCTION structured_coherence(BN,narration,evidenceNodes,evidenceStates)
    parentedNodes <- vector of non-root nodes in BN
    ecsList <- []
    FOR EACH parentedNode IN parentedNodes
        variants <- parents_child_possible_states(parents,parentedNode, narration)
        variants_count <- length(variants)
        sumParentsNarr <- 0
        FOR EACH variant IN variants
            childPrior <- prior probability of the child state in variant
            childPosterior <- posterior probability of the child state in variant,
                obtained by updating on the parents states from this variant
            parentsEvidence[variant] <- joint probability of
                the parents states in variant in BN updated with
                evidenceStates of evidenceNodes
            sumParentsEvidence <- sumParentsEvidence +  parentsEvidence[variant]
            z[variant] <- z_confirmation_measure(childPrior, childPosterior)
        ecs <- 0  #expected confirmation score
        FOR EACH variant IN variants
            IF parentsEvidence[variant] > 0 THEN
                weight <- parentsEvidence[variant]/sumParentsEvidence
            ELSE
                weight <- 1/variants_count
            zScaled <- z[variant] * weight
            ecs <- ecs + zScaled
        ecsList.add(ecs)
    RETURN coherence_from_ecs(ecsList)
```

References

Akiba, K. (2000). Shogenji's probabilistic measure of coherence is incoherent. *Analysis*, *60*(4), 356–359. https://doi.org/10.1093/analys/60.4.356

Allen, R. J. (2010). No plausible alternative to a plausible story of guilt as the rule of decision in criminal cases. In J. Cruz & L. Laudan (Eds.), *Prueba y esandares de prueba en el derecho*. Instituto de Investigaciones Filosoficas-UNAM.

Bovens, L., & Hartmann, S. (2004). *Bayesian epistemology*. Oxford University Press.

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian measures of evidential Support: theoretical and empirical Issues. *Philosophy of Science*, *74*(2), 229–252. https://doi.org/10.1086/520779

Douven, I., & Meijs, W. (2007). Measuring coherence. *Synthese*, *156*(3), 405–425. https://doi.org/10.1007/s11229-006-9131-z

Easwaran, K., & Fitelson, B. (2015). Accuracy, coherence, and evidence. In T. S. Gendler & J. Hawthorne (Eds.), *Oxford studies in epistemology* (Vol. 5). Oxford University Press.

Fenton, N., & Neil, M. (2018). *Risk assessment and decision analysis with bayesian networks*. Chapman;

Hall.
Fitelson, B. (2003). A Probabilistic Theory of Coherence. *Analysis*, *63*(3), 194–199.
Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619
Glass, D. H. (2002). Coherence, Explanation, and Bayesian Networks. In G. Goos, J. Hartmanis, J. van Leeuwen, M. O'Neill, R. F. E. Sutcliffe, C. Ryan, ... N. J. L. Griffith (Eds.), *Artificial Intelligence and Cognitive Science* (Vol. 2464, pp. 177–182). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45750-X_23
Koscholke, J. (2016). Evaluating test cases for probabilistic measures of coherence. *Erkenntnis*, *81*(1), 155–181. https://doi.org/10.1007/s10670-015-9734-1
Meijs, W., & Douven, I. (2007). On the alleged impossibility of coherence. *Synthese*, *157*(3), 347–360. https://doi.org/10.1007/s11229-006-9060-x
Merricks, T. (1995). Warrant entails truth. *Philosophy and Phenomenological Research*, *55*, 841–855.
Olsson, E. J. (2001). Why Coherence Is Not Truth-Conducive. *Analysis*, *61*(3), 236–241.
Olsson, E. J. (2005). The Impossibility of Coherence. *Erkenntnis*, *63*(3), 387–412. https://doi.org/10.1007/s10670-005-4007-z
Pennington, N., & Hastie, R. (1991). A cognitive theory of juror decision making: The story model. *Cardozo Law Review*, *13*, 519–557.
Roche, W. (2013). Coherence and probability: A probabilistic account of coherence. In M. Araszkiewicz & J. Savelka (Eds.), *Coherence: Insights from philosophy, jurisprudence and artificial intelligence* (pp. 59–91). Dordrecht: Springer.
Schippers, M. (2014). *Probabilistic measures of coherence: From adequacy constraints towards pluralism*. *191*(16), 3821–3845. https://doi.org/10.1007/s11229-014-0501-7
Schippers, M., & Koscholke, J. (2019). A General Framework for Probabilistic Measures of Coherence. *Studia Logica*. https://doi.org/10.1007/s11225-019-09848-3
Schupbach, J. N. (2008). On the alleged impossibility of bayesian coherentism. *Philosophical Studies*, *141*(3), 323–331. https://doi.org/10.1007/s11098-007-9176-y
Scutari, M., & Denis, J.-B. (2015). *Bayesian networks in* ***r***. CRC Press.
Shogenji, T. (1999). Is coherence truth-conducive? *Analysis*, *59*(4), 338–345.
Shogenji, T. (2001). Reply to Akiba on the probabilistic measure of coherence. *Analysis*, *61*(2), 147–150. https://doi.org/10.1093/analys/61.2.147
Shogenji, T. (2006). Why does coherence appear truth-conducive? *Synthese*, *157*(3), 361–372. https://doi.org/10.1007/s11229-006-9062-8
Siebel, M. (2004). On Fitelson's measure of coherence. *Analysis*, *64*, 189–190.
Siebel, M. (2006). Against probabilistic measures of coherence. In *Coherence, truth and testimony* (pp. 43–68). Springer.
Spottswood, M. (2013). Bridging the gap between bayesian and story-comparison models of juridical inference. *Law, Probability and Risk*, mgt010.
Vlek, C. (2016). *When stories and numbers meet in court: Constructing and explaining bayesian networks for criminal cases with scenarios*. Rijksuniversiteit Groningen.