An information theoretic analysis of coherence

Rafael Fuchs
Graduate School of Systemic Neurosciences
LMU Munich, Germany

Abstract

The aim of this paper is to import basic concepts from information theory into the epistemological debate about probabilistic measures of coherence. Rather than putting forward and defending a new measure, this paper will sketch an account of 'external' coherence, which will be defined as a relation between a target variable of interest, sources containing (more or less) information about the target, and a rule or theoretical hypothesis that postulates the relevant connections between source and target. Relations with and potential insights for standard notions of coherence in formal epistemology are explored. More generaly, the paper explores the potential benefits of applying information theory to the epistemological debate about coherence.

Keywords: Coherence; Information Theory; Bayesian epistemology; Neurophilosophy; Reasoning; Higher-order cognition

Introduction

Coherence is an important feature of human cognition. Coherent narratives are not only easier to follow, but they also appear to be more convincing than less coherent narratives. Epistemology investigates questions about coherence in relation to truth and epistemic justification. Thus, a central question that motivates the epistemological investigation of coherence concerns whether a strongly coherent belief system or a coherent set of testimonies is more likely to be true than a less coherent one (BonJour, 1985). However, answering this question first requires answer to some more basic questions: what does it even mean to say that a set of statements is "coherent"? How can we measure coherence? These questions are investigated by formal epistemology[1]. Over the last few decades the epistemological literature has generated a large amount of candidate measures of coherence, and their competition goes on. However, the literature on probabilistic measures of coherence has so far neglected insights that can be drawn from information theory and its applications in computational neuroscience and psychology, where modeling the coherent integration of information is a thriving field of research. Quite possibly, these neighbouring fields harbour potent formal tools and valuable empirical results that may help us to get a more well-founded grip on the general questions raised by epistemology: what do we mean by "coherence"? Where does our preference for coherence come from, and (under what conditions) is it a reliable cognitive tool? This paper is intended as a first step towards closing this gap, by applying basic concepts from information theory to build an extended account of coherence. Rather than putting forward and defending a new measure, this paper will sketch an account of target-relative, or *external* coherence, which will be defined as a relation between a target variable T of interest, sources S containing (more or less) information about T, and a rule or theoretical hypothesis Θ that postulates the relevant connections between S and T. We will consider the conditional mutual information between S and T given Θ as a preliminary proposal for quantifying external coherence. However, at this point nothing hinges on this proposal, as the goal of this contribution is only to (i) sketch the basic elements of this alternative notion of coherence, in relation to the standard notion of internal coherence, and to (ii) demonstrate how the debate can benefit from applying information theoretic concepts in general. We will see that investigating external coherence can provide valuable insights for the broader study of coherence, and possibly also connect to other fields of epistemology. In order to start this project, the next section provides a quick overview of basic information theoretic concepts that have been employed in psychology and computational neuroscience for modeling the integration of several information sources into a coherent whole. This is followed by a brief review of the epistemological project of measuring coherence in terms of probability. These two sections together will finally motivate the formulation of a basic account of external coherence. The paper ends with a brief conclusion.

Uses of Information Theory in Neuroscience

Psychologists and neuroscientists have studied the question of how the brain integrates information into a coherent whole in several respects. Early debates occurred between Gestalt psychologists, who emphasised the importance of the holistic perception of an environment, and more atomistic approaches to perception (Wagemans et al., 2012). Later on, the development of information theory led to the formulation of more precise principles, which are grounded in the predictability of the perceptive stimulus (Attneave, 1954; Kareev, 2000; Van der Helm, 2000). More recently, in the context of predictive coding and the free energy principle, visual representations are considered as hypotheses that encode diverse impressions, from which expectations about future

[1] for a discussion of philosophical foundations, related to the Bayesian approach to epistemology see (Bovens & Hartmann, 2004), in particular chapter 2; for overviews of probabilistic measures of coherence, see (Roche, 2013).

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

states are derived. Actions – like saccadic eye movements – are interpreted as experiments to test the respective hypothesis (Friston, Adams, Perrinet, & Breakspear, 2012). In other words, diverse perceptual inputs are subsumed under a unifying cognitive model, which generates a coherent representation of the world. Let us now see how the coherence of perceptual experiences in the light of a cognitive model can be described in terms of information theory.

The general idea is that a representation is coherent, if several parts can be predicted from other parts. For example, if a fraction of a picture is known, and we increase the portion that is known, the rest can be reconstructed with increasing ease. This means that coherent representations can be compressed, and therefore they allow for an efficient reconstruction and prediction of details from a few key characteristics, in which information is concentrated. For example, in a visual field, certain areas have a higher density of information, whereas others follow the same regularity, and hence, the brain can focus on the characteristic points that contain more information, in order to derive a representation of the whole picture (Attneave, 1954).

Furthermore, this kind of redundance also makes information transmission robust against errors, which is an important feature of natural language. Human languages are a mix of predictability and surprise – and this is beneficial, because the predictability (redundance) increases robustness (e.g. against external noise that distorts the original message), while the flexible, non-determined components allow for the expression of new ideas. The statistical structure of natural language, which can be modeled with Markov chains that encode different levels of conditional dependence was first demonstrated by Claude Shannon, in his seminal work (Shannon, 1948).

In computational neuroscience, information theory is used to model communication within neural networks. In particular, the concept of mutual information is often used as a foundation for more complex hypotheses about the computational goals of the brain (Kay & Phillips, 2011; Gutknecht, Wibral, & Makkeh, 2021; Williams & Beer, 2010). The mutual information between two random variables X, Y is standardly defined as

$$I(X;Y) := H(X) - H(X|Y), \tag{1}$$

which is equivalent to:

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \tag{2}$$

where $H(X)$ is the Shannon entropy of the random variable X, $H(X|Y)$ is the conditional entropy of X given Y, and $H(X,Y)$ is the joint entropy of X and Y.

It is important to mention that information theoretic measures, like Shannon entropy, have two components: (i) the measure of information, which is defined pointwise (i.e. the information of a single event), and (ii) the average information contained in a whole probability distribution, which is the expected value of pointwise information under the given probability distribution. The pointwise information[2] $h(\mathbf{X} = \mathbf{x}_i)$ of a single event $\mathbf{X} = \mathbf{x}_i$ (i.e. a specification of all the variables contained in a random vector) is given by

$$h(\mathbf{X} = \mathbf{x}_i) = \log \frac{1}{p(\mathbf{X} = \mathbf{x}_i)}, \tag{3}$$

and the entropy $H(\mathbf{X})$ of the whole ensemble is the expected value, i.e. in the discrete case:

$$H(\mathbf{X}) = \sum_{i=1}^{n} p(\mathbf{X} = \mathbf{x}_i) \log \frac{1}{p(\mathbf{X} = \mathbf{x}_i)}. \tag{4}$$

The mutual information between X and Y measures how informative the determination of one variable is with respect to the possible values of the other variable. As can be seen from equation 2, this relation is symmetric. In computational neuroscience, mutual information is used to measure how much information a *source* (say, the firing patterns of an input neuron or a receptor) provides to the receiver (a processing neuron) about the value of a *target* variable (the object that provided the stimulus). This can also be applied to the systemic level, where sources are external stimuli, and the target is an adequate representation of the object that caused the stimuli.

Turning to epistemology now, here we are concerned with higher-level representations (beliefs and propositions), and in particular, whether their coherence is associated with truth. Since the framework of predictive coding exhibits parallels to the scientific process (e.g. subsumption of sensory data under a hypothesis, and action as experiments to test the hypothesis), it is plausible that applying insights from one field to another can generate mutual benefits. Thus, let us now briefly review the project of developing probabilistic measures of coherence in epistemology.

Probabilistic Measures of Coherence

The coherence of a set of propositions (generally called *information set*) is understood as the degree to which the propositions in this set 'fit together' or 'hang together'. If there is a tension among some propositions – in the extreme case: if some of the propositions logically exclude each other – the set is considered incoherent; if, on the other hand, the propositions 'fit together' – in some sense to be specified – the set is considered to be coherent. The challenge is how to expain what it means to "fit together", how this can be measured, and what the epistemic value of coherence is – in particular, to what extent it is truth conducive.

Probabilistic measures of coherence abstract from the question of what *makes* a set of propositions coherent, and instead focus on measuring coherence as a function of probabilities. If probabilities are also the basis for measures of

[2] I use *ITALIC CAPITAL* letters for random variables, and *italic small* letters for value assignments (instantiations), **BOLD CAPITAL** letters for vectors of random variables, and **bold small** letters for their instantiations.

explanatory power, confirmation, and other relations between propositions (which make them coherent), then a probabilistic measure of coherence also captures the central aspects of coherence. The task for such a measure is then to explicate the notion of 'fitting together' in probabilistic terms.

There are at least three general classes of probabilistic measures of coherence (allowing for intermediates and mixtures), each of which has generated several candidates (Shogenji, 1999; Olsson, 2002; Fitelson, 2003; Douven & Meijs, 2007; Roche, 2013): coherence as correlation, coherence as relative overlap, and coherence as average mutual support or confirmation. For illustrative purposes, let us have a quick look at two exemplary competitors: Shogenji's and Olsson's measures.

Coherence as correlation: the Shogenji measure

One of the earliest measures commonly discussed in the literature is Shogenji's (1999) measure, which defines coherence as correlation, or deviation from independence. Given a set Γ, $C_S(\Gamma)$ is given by:

$$C_S(\Gamma) = \frac{P(\bigwedge_{A\in\Gamma} A)}{\prod_{A\in\Gamma} P(A)}. \quad (5)$$

The neutral point (neither coherent nor incoherent) is given whenever Γ is independent, i.e. $P(\bigwedge_{A\in\Gamma} A) = \prod_{A\in\Gamma} P(A)$, and thus $C_S(\Gamma) = 1$. If $C_S(\Gamma) > 1$, there is net positive dependence among the elements of Γ (hence, Γ is coherent), and if $C_S(\Gamma) < 1$, there is net negative dependence (Γ is incoherent)[3].

If C_S is defined (i.e. $P(A) > 0$ for all $A \in \Gamma$), then C_S is minimal, whenever $P(\bigwedge_{A\in\Gamma} A) = 0$, and maximal whenever $P(\bigwedge_{A\in\Gamma} A) = \min_{A\in\Gamma} P(A)$. Note that this maximum is always relative to the individual probabilities $P(A)$ – in particular, $P(\bigwedge_{A\in\Gamma} A) = 1$ entails $P(A) = 1$ for all $A \in \Gamma$, and thereby $C_S(\Gamma) = 1$.

Coherence as relative overlap: the Olsson Glass measure

The first candidate of this class was independently proposed by Olsson (Olsson, 2002) and Glass (Glass, 2002):

$$C_O(\Gamma) = \frac{P(\bigwedge_{A\in\Gamma} A)}{P(\bigvee_{A\in\Gamma} A)} \quad (6)$$

The idea is that a set of propositions becomes more coherent, if its conjunction covers an increasing area of its disjunction (or set theoretic union). That is, if at least one of the propositions in Γ being true makes it very likely that all of Γ is true, then Γ is highly coherent[4].

As for C_S, C_O is maximal ($C_O = 1$), whenever all propositions in Γ are equivalent, and minimal ($C_O = 0$), whenever they cannot be true together (i.e. the probability of the conjunction is zero). However, unlike C_S, the maximum for C_O is always fixed to 1, hence it is *not* relative to the indivudal $P(A)$ in Γ; so, in particular, if $P(\bigwedge A) = 1$, then also $C_O(\Gamma) = 1$. Furthermore, C_O doesn't have a neutral point, unlike C_S.

Internal vs external coherence

As mentioned previously, the goal of these competing measures is to capture the idea of propositions 'fitting together'. In any case, this fitting together is something beyond the probability of the propositions just being *true* together – which is given by the probability of their conjunction (i.e. their joint probability). It is easy to see that for both, C_S and C_O it is possible that a set of statements with a lower joint probability can be more coherent than a set of statements with a higher joint probability. Thus, we may call this kind of coherence – how well propositions fit together – the *internal* coherence of an information set. In the next section, this internal coherence will be distinguished from the *external* coherence (of an information set).

In the literature on probabilistic measures of coherence, intuitive counterexamples are offered against most, if not all, serious candidate measures (Fitelson, 2003; Bovens & Hartmann, 2004; Douven & Meijs, 2007). However, often enough individual intuitions about examples diverge. An alternative approach for singling out the best candidates is the formulation of general normative principles, which an adequate measure has to satisfy. However, some principles that received strong intuitive support turned out to be jointly inconsistent (Schippers, 2014). So, in order to make progress on that front, the literature must seek principled and independent reasons to justify the precedence of one principle over another one. Empirical investigations, like (Koscholke & Jekel, 2017; Harris & Hahn, 2009) may also help to shed further light on our intuitions regarding internal coherence. However, in order to produce deeper insights, empirical research needs to be tied to theoretical insights, e.g. from computational neuroscience, which can help to explain and predict our intuitions about coherence in different domains, in order to meaningfully compare them to the normative requirements produced by the epistemological study of coherence and its relation to truth.

In any case, there is an important problem, which arguably applies to all probabilistic measures of internal coherence: the problem is posed by propositions about which we are extremely confident (i.e. their marginal and joint probabilities are high), but at the same time they are considered to be unrelated. In the extreme case, the information set consists only of *known facts* (which we take to have probability 1), and thereby they automatically become probabilistically independent of everything else. For the Olsson measure, if the joint probability is 1, we always get maximal coherence, regardless of the theoretical relations between the propositions involved. On the other hand, the Shogenji measure is always neutral for maximal joint probability. For average mutual support measures, which are based on measures of

[3] Note that this measure is exactly proportional to a single pointwise contribution in Watanabe's ((1960, 1961)) measure of total correlation, which is also a non-negative generalisation of mutual information for more than two random vectors.

[4] note that $C_O(\Gamma) = P(\bigwedge\Gamma|\bigvee\Gamma)$.

confirmation, this is just an instance of the problem of old evidence[5] (Glymour, 1980). However, our problem extends beyond propositions with maximal probability: if we consider propositions in which we are highly confident, but they are considered as unrelated, several coherence measures will output a high degree of coherence. For example, the set of propositions $\Omega = \{$ C_1: There will be a presidential election in the US in 2024; C_2: In Munich, there will be at least one day with a maximum temperature above 25 degrees Celsius in the summer of 2023$\}$ turns out to be highly coherent under the Olsson- and several average mutual support measures, because we have extremely high confidence in both propositions (say, $P(C_1) = P(C_2) = 0.99$, then $C_O(\Omega) \approx 0.98$). But obviously, these are completely unrelated. Shogenji's measure does better, as it outputs a neutral value for independent propositions. However, for known propositions (with probability 1), C_S will *always* output a neutral value, which is too strict. After all, in scientific and legal practice, we often use one set of known facts (together with a theory) to explain another set of facts.

This problem is not necessarily lethal, but intuitions regarding its implications may again pull in different directions: should certain propositions always be considered neutral or always maximally coherent? Should extreme probabilities be excluded from the analysis of coherence? Or should there be some additional factor that informs measures of coherence, beyond their actual probability? In this case, there has to be a case-wise definition of the respective measure (e.g. for propositions with probability 1, assign degree of coherence x if some designated property holds, and assign y otherwise). However, in order to do this in a non-arbitrary way, we will need to discuss additional theoretical principles that go beyond internal coherence. Achieving this is a potential contribution of the notion of external coherence.

Thus, the plan for the remainder of this paper is to explain in more detail what is meant by external coherence, how it is different from internal coherence, and how it can be a useful concept for the epistemological study of coherence.

Towards an account of external coherence

Now it is time to explain what is meant by external coherence. The general idea is that a set of information sources (e.g. testimonial statements) is externally coherent to the extent that it is jointly informative about a target variable of interest. This idea is motivated by its analogy to the integration of perceptual inputs into a coherent representation, as discussed in the beginning of the paper. In the context of high-level cognition, organisational principles that connect individual thoughts into a unified scenario seem to play a role analogous to binding principles (hypotheses) in perception. This becomes particularly apparent when assessing a set of known facts: are these facts connected? Do they provide any information about related questions, or do they appear to be completely unrelated? It seems that these factors affect our judgement about how well such known facts fit into a coherent sequence of thoughts (argument or narrative) that is constructed in the service of finding answers to a specific set of related questions.

Targets, sources, and theories

To make this general idea more precise, we need to formally introduce the notion of a target system, which consists of source variables and target variables. The source variables are pieces of information that an agent obtains from the target system. Thus, a set of *instantiated* source variables is an *information set*, in the epistemological sense. The target variables represent the questions on which our investigation focuses. It is important to note that the classification of a variable as a source- or target variable is context-dependent: e.g. a proposition can be a source variable in one context, and a target variable in another context. However, the nature of the concrete problem will make the assignment of source and target (i.e. what we already know beforehand and what we want to know in the end) obvious. The connection between source- and target variables is established by a *theory*, which is explained right below. The formal notation will be as follows:

1. The *target variables* are represented as a vector of random variables $\mathbf{T} = \langle T_1, ..., T_m \rangle$. They are the central questions of interest.

2. The *source variables* are represented by another vector of random variables $\mathbf{S} = \langle S_1, ..., S_n \rangle$.

3. The theory, denoted as Θ, *connects* the source- and target variables, and it can be taken as the basis for assigning probabilities and postulating conditional dependencies. Formally, Θ is a *set of constraints* that any probability function must satisfy to be consistent with the theory (an example, to be presented right below, will make this more intuitive). This set of constraints can be subdivided into several subsets. For our purpose, two categories will be sufficient: let $\Theta_i = \langle \vartheta_i, \alpha_i \rangle$, where

   ϑ_i represents the core hypothesis, or -principles. These are *general* statements, like scientific laws or rules that apply to any system in the domain of the theory, and they specify relations between variables within the system. These relations can be deterministic in principle (like laws in classical mechanics), but they can also be probabilistic, if the most general description of the system is only stastical (e.g. due to complex interactions). In the simplest case, when we are dealing with propositional variables, ϑ_i specifies conditional probabilities $p(X|Y)$ between variables X, Y.

   α_i represents *particular knowledge* about variables in the target system, like observations or assumed boundary conditions. When dealing with propositional variables, α specifies unconditional probabilities.

[5] The problem of old evidence (Glymour, 1980) consists in the fact that, by Bayes' theorem, $P(E) = 1$ entails that $P(H|E) = P(H)$, i.e. already known facts apparently cannot confirm a hypothesis, even if the hypothesis is able to explain those facts.

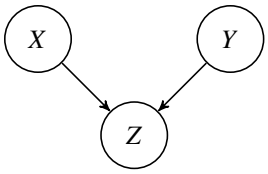

Figure 1: Causal DAG for the car example. The general rules in ϑ determine the conditional independence structure (which is represented in the graph), and together with particular knowledge contained in α, the graph can be equipped with a probability distribution, to specify a Bayesian network.

With all three parameters at hand, the current proposal is to identify external coherence $\mathcal{C}$ with the conditional mutual information between $\mathbf{S}$ and $\mathbf{T}$, given a specific theory, Θ_i. Formally,

$$\mathcal{C}(\mathbf{S};\mathbf{T};\Theta_i) := I(\mathbf{S};\mathbf{T}|\Theta_i). \tag{7}$$

Now, let us look at an example – adapted from (Sprenger & Hartmann, 2019, p. 194) – to develop an intuitive understanding of this definition. Suppose you want to know whether the engine of your car might break down during your trip (variable Z), and you know that there are just two critical components[6] that can cause the engine to fail: a dead battery (X) or a blocked fuel pump (Y). Thus, in this situation, Z is our target variable, and X, Y are the source variables. Now, what is our background theory Θ? As for the general rules ϑ, suppose you know that the engine will work ($Z = z$) *if and only if* both, the battery is not empty ($X = \neg x$) and the pump works ($Y = y$). Furthermore, X and Y are independent (notation: $X \perp Y$). So, in our example, these two rules ($z \leftrightarrow \neg x \wedge y$ and $X \top Y$) are the only relevant rules, i.e. $\vartheta = \{z \leftrightarrow \neg x \wedge y, X \perp Y\}$. The relevant causal structure can be represented as a DAG (fig. 1). Furthermore, we can assume that you also have some more specific information about the components in the target system. That is, α contains information about the state of some of the variables in the system. For example, suppose that you went to a car repair shop recently, and so you know that battery and pump are in optimal condition — however, since there is also a bit of noise in the system (any kind of problem that might come up during your trip and that you can't exactly predict at the current moment), your degrees of belief are $p(\neg x) = 0.99$ and $p(y) = 0.9$ (i.e. α contains these two unconditional probabilities). Together with $\vartheta = \{z \leftrightarrow \neg x \wedge y, X \perp Y\}$, this entails that $p(z) = 0.9 \cdot 0.99 = 0.891$.

The explicit reference to a background theory Θ, where general rules θ and boundary conditions α are separated, can help to overcome the difficulties with coherence relations among known facts. In a nutshell, the idea is to vary α, in order to show how counterfactual initial conditions (such that the known facts are assumed to be not yet instantiated) affect the joint informativeness of the sources regarding the target system. The resulting mutual information under counterfactual initial conditions then measures how informative the sources can be in principle (under a fixed core hypothesis ϑ) with respect to the target. This is broadly in line with counterfactual causal explanations in the philosophy of science (Woodward & Hitchcock, 2003).

Finally, if there is more than one potential target variable, there is an important difference between external coherence, relative to a target variable, and the internal coherence among the source variables (i.e. the information set). Consider another standard example (Bovens & Hartmann, 2004, p. 29): $\{a_1$: My pet Tweety is a bird; a_2: My pet Tweety cannot fly$\}$. With adequate probabilistic assumptions (most birds can fly), our standard measures of internal coherence, such as C_S, C_O, and several confirmation-based measures will assign a rather low value to $\{a_1, a_2\}$ – after all, there is some tension in the set. However, depending on what the target question is, the *external* coherence of that set becomes orthogonal to its internal coherence. For example, if the goal is to know whether Tweety is a penguin, a_1 and a_2 together become highly informative. On the other hand, if the question is whether I got this pet from my parents, both pieces of information are irrelevant, given commonsense background knowledge (which can of course change if we obtain more specific knowledge about my own or my parents' preferences for pets). If the information *is* informative (e.g. relative to the question whether Tweety is a penguin), there are two ways in which it can be informative: if we want to know whether the question can confirm or rule out a specific answer to our question (Tweety is/isn't a penguin), we are looking at the *pointwise* contribution to mutual information. Otherwise, if we want to know how informative *any* information from the source variables A_1, A_2 can be with respect to obtaining *any* answer to the target question (whether Tweety is a penguin), we are looking at the expected information gain, relative to the full probability distribution. This point brings us to the more general contribution that information theory cann offer to the study of coherence, which we will now outline in the short remainder of this section.

The contribution of information theory to understanding coherence

The Tweety-example pointed to an important novelty in using an information theoretic measure for external coherence: it means that we can talk about the coherence of *topics* or *questions*, rather than only particular *statements*. This is, because the measure is an expected value over a whole probability distribution, containing all point-probabilities for instantiated propositions. This makes an information theoretic approach broader than the standard probabilistic approach to internal coherence: instead of asking how well a set of *statements* fits together, we can ask how informative any answer

[6] in principle, of course, we could make this scenario arbitrarily complex, but for the purpose of developing an intuitive understanding this simple example should be sufficient.

to the *questions* encoded by the random variables will be, on expectation. For example, in the information set Ω from above, we can map the positive statements c_1, c_2 to binary random variables C_1, C_2 (their possible values represent the positive and negative literals $c_i, \neg c_i$, for $i = 1, 2$ respectively). Then, we can ask what a positive or negative answer to the question whether there will be a presidential election in 2024 (i.e. $C_1 = c_1$ or $C_1 = \neg c_1$) may tell us about the question whether there will be at least one day with a temperature above 25 degrees in Munich, in summer 2023 (variable C_2). The probability distribution then represents our expectation to receive the respective answers (positive or negative). Since the weather in Munich is independent from the US presidential relations, learning the value of one variable won't tell us anything about the other variable, and hence, their mutual information is zero.

Furthermore, information theory can help more generally with the further development of probabilistic measures of coherence. For example, analysing mutual information can explain some cases that were previously considered as counterexamples to the Shogenji measure. Consider the following two scenarios (Douven & Meijs, 2007, p. 414). In scenario 1, we are investigating a murder that occurred in a big city with 10,000,000 inhabitants. 1,059 inhabitants are Japanese, and 1,059 own a Samurai sword. Of those, 9 are both, Japanese and owners of a Samurai sword. We want to find the murderer among the 10,000,000. In scenario 2, we already know that the murder occurred in a particular street, where 100 people live. 10 of those are Japanese, 10 have a Samurai sword, and 9 are Japanese and own a Samurai sword. Now, we are asked to assess the coherence of the following information set: $\{j$: the suspect is Japanese; o : the suspect owns a Samurai sword$\}$.

Douven & Meijs argue that intuitively, the information set $\{j, s\}$ is more coherent in scenario 2 than in scenario 1. This is also the verdict of several probabilistic measures of coherence that were tested by Douven & Meijs (2007, p. 414). However, for the Shogenji measure, in scenario 1 we obtain $C_S(j, o) = 80.3$, whereas in scenario 2 we obtain $C_S(j, o) = 9.0$, which seems to be the wrong result. Now note that for two variable-instantiations x, y $C_S(x, y)$ is proportional to their pointwise mutual information, and thus, we can use our information theoretic framework to explain the above result. Let us look back to equation 1, which defines the mutual information for random variables X, Y as the difference $H(X) - H(X|Y)$. We can also do this for a single point, writing $h(x) - h(x|y) = \log p(x|y) - \log p(x)$. The interpretation of this formula is as follows. It tells us by how much learning y before learning x *reduces* the surprise of learning x afterwards. It is maximal, when $P(x|y) = 1$, i.e. x is entailed by y. In this case, $h(x|y) = 0$, and thus, the difference is $h(x)$, which means that the amount by which the surprise upon learning x is reduced after learning y is exactly the amount of surprise (information) that lies in the event x before learning anything. Hence, pointwise mutual information (and by extension, the Shogenji measure) is actually a measure that is *relative* to the (pointwise) information in the whole ensemble. If $h(x)$ is very high (because $p(x)$ is very low), then a slight reduction of the surprise of learning x after learning y can be greater than a large reduction in the case where $h(x)$ is already comparatively low. This is precisely what happens in Douven & Meijs' example: in scenario 1, the marginal probabilities of o and j are extremely low, whereas the corresponding probabilities in scenario 2 are *comparatively* much higher. But how could we fix this? It is possible to normalise average mutual information (and variants of it), and one can also normalise the pointwise measure, relative to the maximum amount of information (surprise) contained in the single and joint variables. Thus, a normalised version of the logarithmised Shogenji measure (or, equivalently, normalised pointwise mutual information) can be defined as

$$NLC_S(x, y) := \frac{h(x) - h(x|y)}{h(x, y)}, \tag{8}$$

where $h(\mathbf{x}) = \log \frac{1}{p(\mathbf{x})}$ (this also gives the 'intuitively correct' result in the preceding example). Whenever X and Y are independent, it follows that $NLC_S(x, y) = 0$ (new neutral value, due to $\log(1) = 0$). Furthermore, for $p(x|y) = 1$, $NLC_S(x, y) = 1$, which is the maximum value, and for $p(x|y) \to 0$, $NLC_S(x, y) \to -1$. For sets containing falsehoods or contradictions ($p(x) = 0$ or $p(x, y) = 0$), $NLC_S(x, y)$ can be defined via the limit as -1.

Conclusion

This paper has developed a basic account of external coherence, as a tripartite relation between a target system T, a set of sources S, and a theory Θ that provides the expected relation between target system and sources. The initial proposal for quantifying this relation is in terms of the conditional mutual information of S and T, given Θ. An important advantage of employing information theory is that we can broaden the focus of coherence considerations from statements to topics or questions, which is given by the expected value over the set of random variables. Furthermore, the explicit reference to a theory Θ allows us to better deal with the coherence of a set of known facts (with probability 1), by varying the state of affairs encoded in the auxiliary assumption of Θ, to allow for counterfactual experiments.

References

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological review*, *61*(3), 183.

BonJour, L. (1985). *The structure of empirical knowledge*. Harvard University Press.

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

Douven, I., & Meijs, W. (2007). Measuring coherence. *Synthese*, *156*, 405–425.

Fitelson, B. (2003). A probabilistic theory of coherence. *Analysis*, *63*(3), 194–199.

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Frontiers in psychology*, *3*, 151.

Glass, D. H. (2002). Coherence, explanation, and bayesian networks. In *Artificial intelligence and cognitive science: 13th irish conference, aics 2002 limerick, ireland, september 12–13, 2002 proceedings 13* (pp. 177–182).

Glymour, C. N. (1980). *Theory and evidence*. Princeton University Press, Princeton.

Gutknecht, A. J., Wibral, M., & Makkeh, A. (2021). Bits and pieces: Understanding information decomposition from part-whole relationships and formal logic. *Proceedings of the Royal Society A*, *477*(2251), 20210110.

Harris, A. J., & Hahn, U. (2009). Bayesian rationality in evaluating multiple testimonies: Incorporating the role of coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(5), 1366.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological review*, *107*(2), 397.

Kay, J., & Phillips, W. (2011). Coherent infomax as a computational goal for neural systems. *Bulletin of mathematical biology*, *73*(2).

Koscholke, J., & Jekel, M. (2017). Probabilistic coherence measures: a psychological study of coherence assessment. *Synthese*, *194*, 1303–1322.

Olsson, E. J. (2002). What is the problem of coherence and truth? *The Journal of Philosophy*, *99*(5), 246–272.

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). Springer.

Schippers, M. (2014). Probabilistic measures of coherence: From adequacy constraints towards pluralism. *Synthese*, *191*(16), 3821–3845.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Shogenji, T. (1999). Is coherence truth conducive? *Analysis*, *59*(4), 338–345.

Sprenger, J., & Hartmann, S. (2019). *Bayesian philosophy of science*. Oxford University Press.

Van der Helm, P. A. (2000). Simplicity versus likelihood in visual perception: From surprisals to precisals. *Psychological Bulletin*, *126*(5), 770.

Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., Van der Helm, P. A., & Van Leeuwen, C. (2012). A century of gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological bulletin*, *138*(6), 1218.

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM Journal of research and development*, *4*(1), 66–82.

Watanabe, S. (1961). A note on the formation of concept and of association by information-theoretical correlation analysis. *Information and Control*, *4*(2-3), 291–296.

Williams, P. L., & Beer, R. D. (2010). Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Woodward, J., & Hitchcock, C. (2003). Explanatory generalizations, part i: A counterfactual account. *Noûs*, *37*(1), 1–24.