Schweizer, Mark

Working Paper

Comparing holistic and atomistic evaluation of evidence

Preprints of the Max Planck Institute for Research on Collective Goods, No. 2012/21

Provided in Cooperation with:
Max Planck Institute for Research on Collective Goods

Suggested Citation: Schweizer, Mark (2012) : Comparing holistic and atomistic evaluation of evidence, Preprints of the Max Planck Institute for Research on Collective Goods, No. 2012/21, Max Planck Institute for Research on Collective Goods, Bonn

This Version is available at:
https://hdl.handle.net/10419/85004

**Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2012/21**

Comparing Holistic and Atomistic Evaluation of Evidence

Mark Schweizer

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2012/21

Comparing Holistic and Atomistic Evaluation of Evidence

Mark Schweizer

November 2012

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Comparing Holistic and Atomistic Evaluation of Evidence

Mark Schweizer*

Abstract

Fact finders in legal trials often need to evaluate a mass of weak, contradictory and ambiguous evidence. There are two general ways to accomplish this task: by holistically forming a coherent mental representation of the case, or by atomistically assessing the probative value of each item of evidence and integrating the values according to an algorithm. Parallel constraint satisfaction (PCS) models of cognitive coherence posit that a coherent mental representation is created by discounting contradicting evidence, inflating supporting evidence and interpreting ambivalent evidence in a way coherent with the emerging decision. This leads to inflated support for whichever hypothesis the fact finder accepts as true. Using a Bayesian network to model the direct dependencies between the evidence, the intermediate hypotheses and the main hypothesis, parameterised with (conditional) subjective probabilities elicited from the subjects, I demonstrate experimentally how an atomistic evaluation of evidence leads to a convergence of the computed posterior degrees of belief in the guilt of the defendant of those who convict and those who acquit. The atomistic evaluation preserves the inherent uncertainty that largely disappears in a holistic evaluation. Since the fact finders' posterior degree of belief in the guilt of the defendant is the relevant standard of proof in many legal systems, this result implies that using an atomistic evaluation of evidence, the threshold level of posterior belief in guilt required for a conviction may often not be reached.

* Max Planck Institute for Research on Collective Goods, Bonn

I. Introduction

In legal trials, fact finders need to form a conviction regarding the truth of factual statements based on a mass of often incomplete, ambivalent and contradicting evidence. There are two fundamentally different ways the fact finder can go about this difficult task: she can either assess the evidentiary strength of each item of evidence and then integrate her individual assessments according to some general rule to arrive at a conclusion, or she can assess the whole mass of evidence globally, forming a holistic overall impression of the case. The former method is sometimes referred to as "atomistic" evaluation of evidence, while the latter is called "holistic" (Twining 2006, p. 309). "Atomistic" and "holistic" can only describe the fact finding process at a very general level; a number of different approaches to the evaluation of evidence fall into each category. In this article, one currently popular model of holistic evaluation of evidence based on cognitive coherence is contrasted with a leading theory of atomistic evaluation of evidence, namely subjective probability theory.

Holistic evaluation of evidence assumes that legal decision making is based on constructing and evaluating coherent interpretations or stories from the available items of evidence (see Pennington & Hastie 1992 for a classic approach). Cognitive coherence theories understand evaluation of evidence as a process of forming a coherent mental representation of the evidence, integrating it with the background knowledge of the subject (Simon *et al.* 2004). A more coherent mental representation leads to higher subjective confidence that the representation is correct (Glöckner *et al.* 2010, p. 219). During the evaluation of the evidence, coherence is maximized by discounting contradicting evidence, inflating supporting evidence and interpreting ambivalent evidence in a way that is coherent with the emerging decision (Simon 2004, p. 522). One important empirical prediction of cognitive coherence theories of evidence evaluation is that the mental model of the case "shifts" during the decision process towards an interpretation coherent with the emerging decision (Holyoak & Simon 1999; Carlson & Russo 2001; Engel & Glöckner 2012). The result of this process, referred to as "coherence shift", is that even when the evidence has little probative weight, the fact finder has a high degree of confidence in having made the correct decision (Simon *et al.* 2004, p. 819). If the standard of proof that has to be met before a fact finder may accept a factual proposition as true is understood as a degree of conviction, or belief, in the truth of the allegation, cognitive coherence theories of evidence evaluation imply that the threshold value may be reached even when the evidence is ambivalent, weak or partially missing (Simon 2004, p. 519).

Interestingly, subjective probability theory is also based on a notion of coherence, but on an entirely different concept of coherence. According to subjective probability theory, the partial beliefs of a subject are coherent if they do not violate the axioms of probability theory, namely positivity, certainty and additivity (Finetti 1937). The subject is assumed to hold some prior belief in the truth of a proposition, which he updates when he learns of new evidence. If the subject is to remain coherent in the sense of subjective probability theory, the updating must be done according to Bayes rule, which is why people who believe that degrees of belief should conform to the axioms of probability theory are often referred to as "Bayesians" (they

might more fittingly be called "coherentists"). Subjective probability theory is primarily a normative theory of forming a conviction in the truth of a proposition; nobody claims (anymore) that it accurately describes the actual psychological process of belief formation (Kaye 1988, p. 178, but see Lagnado 2011). Whether it is applicable in the context of evidence evaluation by judicial fact finders is subject to a decades old controversy (see Tillers 2011 for an overview). Ensuring the coherence of partial beliefs, in the sense of subjective probability theory, quickly becomes computationally intractable (Callen 1982). However, in the late 1980s, algorithms for inference in so called "Bayesian networks" were developed, which allow the compact representation of the full joint probability distribution using a directed graph and conditional probabilities (Pearl 1988). A number of authors have suggested using Bayesian inference networks for the evaluation of evidence in legal contexts (Edwards 1991; Robertson & Vignaux 1992; Kadane & Schum 1996; Taroni *et al.* 2006; Fenton & Neil 2011; Juchli *et al.* 2012).

Both cognitive coherence theories and subjective probability theory are models of belief formation. Both can be understood as models of causal inference (Thagard 2004). This paper investigates how the posterior belief in the truth of a main hypothesis, in this case whether the defendant is guilty of taking money from a safe, differs when the evidence is evaluated holistically versus atomistically. For the atomistic evaluation of the evidence, the prior beliefs and the likelihoods for each item of evidence and each intermediate hypothesis are elicited from the subjects. The resulting parameters are then integrated using a Bayesian network, allowing the computation of the posterior belief in the truth of the main hypothesis for each subject based on her own partial beliefs. This posterior degree of belief is the degree of belief the subject should have, provided her partial beliefs are coherent in the sense of subjective probability theory, and can be contrasted with the degree of belief in the guilt of the defendant based on a holistic assessment of the case. The main result of the experiment reported in this paper is that the average degree of belief in the guilt of the defendant of those who convict is *lower* when the evidence is assessed atomistically versus holistically, while it is *higher* for those who acquit. While the subjects interpret the same evidence in completely different ways when they assess it holistically, their computed posterior probability of guilt converges when their atomistic assessments are integrated according to the logic of subjective probability theory.

The rest of this article is structured as follows: in Part II the parallel constraint satisfaction model of cognitive coherence is briefly explained, and Part III provides a cursory introduction into subjective probability theory and Bayesian networks. Part IV sets out the hypotheses to be experimentally tested. Part V describes the experiment, its results, and its limitations. The conclusion summarizes the main contributions of this article.

II. Holistic Evaluation of Evidence

A. Coherence Construction by Parallel Constraint Satisfaction

In line with a basic claim from Gestalt psychology, cognitive coherence theories regard the assessment of evidence as holistic and relying at least partially on an automatic process that has been adapted from perception (Simon *et al.* 2004). The process of constructing cognitive coherence can be computationally implemented as a parallel constraint satisfaction (PCS) process (Thagard 1989; Thagard & Verbeurgt 1998; Holyoak & Simon 1999). A constraint is a relationship between two cognitions (propositions). The coherence problem consists of dividing the set of propositions in two sub-sets of accepted (or true) and rejected (or false) propositions in a way that satisfies the most constraints. If two propositions are coherent (fit together), the constraint is positive and it is satisfied if the two statements connected by it are in the same sub-set, while an incoherent relationship is represented by a negative constraint which is satisfied if the two statements connected by it are in different sub-sets (see Thagard 2000, p. 16 seq., for a full exposition). The strength of the (in)coherence is expressed as weight of the constraint. In most cases, not all the constraints can be satisfied at the same time. The goal is divide the propositions into "accepted" and "rejected" propositions so that the weight of the satisfied constraints is maximized.

There can be no general algorithm that exactly solves all parallel constraint satisfaction problems in polynomial time (Thagard & Verbeurgt 1998). However, a number of algorithms for approximate solutions are available; the most popular, and the one almost exclusively used in psychological research, uses a representation of the problem in a connectionist network (Read *et al.* 1997). In a connectionist network, positively linked variables excite each other while negatively linked variables inhibit each other. In an iterative process, activation spreads through the network. Each and every element influences, and is influenced by, the entire network, so that every processing cycle results in a slightly modified state of the network. The core feature of constraint satisfaction mechanisms is that the connectionist network will reconfigure itself until the constraints settle at a point of maximal coherence. This process forces coherence upon a mental representation of the task that is initially incoherent in complex decisions. Since the links between nodes in a connectionist network are bidirectional, the evidence influences the hypotheses, but the activation of the hypotheses also influences the interpretation of the evidence (Holyoak & Simon 1999). The formation of coherence in an iterative process therefore leads to a *polarization of the evidence*: evidence that supports the emerging decision is strongly endorsed while contradicting evidence is dismissed, rejected, or ignored. These so called "coherence shifts" or more generally "predecisional information distortions" (Russo *et al.* 2008) have been demonstrated in a variety of decision making tasks (Brownstein 2003), most notably also for legal decision making (Holyoak & Simon 1999; Carlson & Russo 2001; Hope *et al.* 2004; Lundberg 2004; Simon *et al.* 2004; Glöckner & Engel 2008; Engel & Glöckner 2012).

B. Coherence Shifts lead to Inflated Confidence

The devaluation of contradicting evidence and the inflation of supporting evidence as well as the interpretation of ambiguous elements as supportive for the emerging decision lead to an inflated confidence in having made the correct decision. In other words, although the evidence of the case is objectively weak, as is evidenced by the fact that the decision makers are split over whether the evidence supports a guilty verdict, both those who find the defendant guilty and those who find him innocent are quite confident that they have made the right choice (Holyoak & Simon 1999; Simon *et al.* 2004; Glöckner & Engel 2008). For example, under a scenario describing the case against a person being accused of taking money from a safe, subjects are split over whether to convict or acquit the defendant (Simon *et al.* 2004; Glöckner & Engel 2008; Engel & Glöckner 2012). But the distribution of the confidence levels of the subjects is skewed towards high confidence in both those who convict and those who acquit the defendant (Simon *et al.* 2004, p. 819). Persons who find the defendant guilty express an average posterior degree of belief in guilt of the defendant which is about twice as high as those who find him innocent (roughly 80% versus 40%, see Glöckner & Engel 2008, p. 13). In a holistic evaluation of the evidence, subjective confidence in the truth of the main hypothesis is mostly the result of coherence shifts during the decision making process, and therefore a questionable standard of proof (but see Glöckner & Engel 2008, which shows that raising the standard of proof does have the desired effect of reducing the number of convictions given the same evidence).

Parallel constraint satisfaction models of coherence-based reasoning have been suggested as both descriptive (Holyoak & Simon 1999; Simon *et al.* 2004; Simon 2004; Engel & Glöckner 2012) and normative (Thagard 2004) models of legal decision making. This is not the place to settle the debate over the normative status of parallel constraint satisfaction models for legal decision making (generally positive AMAYA 2008, p. 307). However, the polarization of evidence predicted by PCS models – leading to inflated support for whichever hypothesis the decision maker accepts – casts doubt on the status of PCS models as *normative* models of evidence evaluation.

III. Atomistic Evaluation of Evidence

A. Subjective Probability Theory as a Normative Model for the Evaluation of Evidence

According to the subjective interpretation of probability, probability is a degree of belief (Finetti 1937). Unlike the frequentist interpretation, the subjective interpretation of probability allows to speak intelligibly of the "probability" of a single case (Hacking 2008, p. 136). "Subjectivists" or "Bayesians", or, as I prefer to call them, "coherentists", believe that the partial beliefs of a subject should (normatively) not violate the axioms of probability theory, i.e., positivity (probability is a real number between 0 and infinity), certainty (the probability of a certain event is 1) and additivity (the probability of one of several mutually exclusive events oc-

curring is the sum of their individual probability). From positivity and certainty follows immediately that probabilities are normalized, i.e., bound between 0 and 1. A variety of arguments can be made why degrees of belief should conform to the axioms of probability theory. The least technical one is that unless the beliefs of a subject conform to the axioms of probability theory, the subject can be made the victim of a "Dutch book", a set of bets that incurs him a certain loss, no matter how the state of the world turns out (Finetti 1937; Christensen 2007, p. 116 seq.).

From the axioms of probability theory follows immediately that the conditional probability of A given B is calculated according to Bayes theorem

$$\Pr(\mathrm{A}|\mathrm{B}) = \frac{\Pr(\mathrm{A},\ \mathrm{B})}{\Pr(\mathrm{B})}. \tag{1}$$

The central importance of Bayes' theorem for subjective probability theory stems from the fact that the subject should update her prior belief in A when she learns that B is the case according to this theorem. For Bayesians, Bayes' theorem is a normative *rule* for rational updating of beliefs (Good 1950, p. 61).

Written differently, equation (1) gives the product rule, which allows the calculation of the joint probability of the two events A and B:

$$\Pr(\mathrm{A},\mathrm{B}) = \Pr(\mathrm{A}|\mathrm{B})\Pr(\mathrm{B}) = \Pr(\mathrm{B}|\mathrm{A})\Pr(\mathrm{A}) \tag{2}$$

Assuming that A can take just two states, true and false (a_{true} and a_{false}), Bayes rule can be rewritten as follows in its odds form, using the product rule to calculate the joint probability of $\Pr(a_{true}, \mathrm{B})$,

$$\frac{\Pr(a_{true}|\mathrm{B})}{\Pr(a_{false}|\mathrm{B})} = \frac{\Pr(\mathrm{B}|a_{true})}{\Pr(\mathrm{B}|a_{false})}\frac{\Pr(a_{true})}{\Pr(a_{false})} \tag{3}$$

This form of Bayes rule makes transparent that it is the ratio $\Pr(\mathrm{B}|a_{true})/\Pr(\mathrm{B}|a_{false})$, called the *likelihood ratio*, that determines the degree of change from prior to posterior odds, or from prior to posterior probability. In subjective probability theory, the likelihood ratio is therefore a *measure of evidentiary strength* (Good 1983, p. 132).

The iterative application of the product rule leads to the chain rule, which allows the calculation of the joint probability of any number of events $A_1, \ldots, A_n$:

$$
\begin{aligned}
&\Pr(A_1, \ldots, A_n) \\
&\quad = \Pr(A_n | A_1, \ldots, A_{n-1}) \Pr(A_1, \ldots, A_{n-1}) \\
&\quad = \Pr(A_n | A_1, \ldots, A_{n-1}) \Pr(A_{n-1} | A_1, \ldots, A_{n-2}) \Pr(A_1, \ldots, A_{n-2}) \\
&\quad = \Pr(A_n | A_1, \ldots, A_{n-1}) \Pr(A_{n-1} | A_1, \ldots, A_{n-2}) \Pr(A_1, \ldots, A_{n-2}) \ldots \Pr(A_2 | A_1) \Pr(A_1) \\
&\quad = \prod_{i=1}^{n} \Pr(A_i \mid A_1, \ldots, A_{i-1}). \qquad (4)
\end{aligned}
$$

The law of total probability says that to calculate Pr(B), the probability of B under all possible states of A must be summed. If A can take on just two states, true or false, then Pr(B) is calculated according to

$$\Pr(B) = \sum_{A} \Pr(A, B) = \Pr(B, a_{true}) + \Pr(B, a_{false}). \qquad (5)$$

Whether subjective probability theory is a useful model for the formation of a belief in the context of the forensic evaluation of evidence is subject to a debate that has been likened to a 40 year war (Park *et al.* 2010, 1) and has been reinvigorated by the decision of the Appeal Court of England and Wales in *R v. T* ([2010], EWCA Crim 2439; for an introduction to the latest round of the controversy see Aitken 2012). Some people take issue with the betting paradigm of subjective probability theory (Cohen 1977, p. 90), while others are convinced that the expression of degrees of belief that are not grounded in observed relative frequencies in mathematical terms will lead to "wholly inaccurate, and misleadingly precise, conclusions" (Tribe 1971, p. 1359; this is essentially also the position of the Appeal Court in *R v. T*). As Taroni et al. have put it, the proof of the pudding is in the eating – the demonstration of the practical use of Bayesian inference should convince sceptics (Taroni *et al.* 2006, p. 23).

B. Bayesian Networks as Decision Aids for the Evaluation of Evidence

Holding partial beliefs that are coherent in the sense of subjective probability theory quickly becomes impossible without some sort of decision aid (Charniak 1991, p. 55). Bayesian networks, also referred to as "belief nets" (Darwiche 2009, p. 71), are a graphical representation of the direct dependencies among a set of variables and force coherence in the sense of subjective probability theory on the set of partial beliefs represented by the network (Charniak 1991, p. 55).

A Bayesian network is a directed acyclic graph in which a node (variable) is connected by a directed edge to another node if the variable represented by the node has a direct influence on the other variable (for a general introduction into Bayesian networks see Taroni *et al.* 2006, p. 33 seq.) A directed graph is *acyclic* if there is no way to start at some node A and follow a sequence of edges that leads back to node A (colloquially, it does not contain a "feedback cycle", Jensen & Nielsen 2007, p. 34). A conditional probability table is associated with each node (root nodes are only associated with "unconditional" or "prior" probabilities), which gives the probability for each mutually exclusive state of the variable given its parents (a par-

ent of a node is an immediate ancestor of this node, i.e., any node that is *directly* connected to the node). In the network used here, each variable can take on only two states which can be interpreted as "true" and "false". Using the concept of conditional independence, Bayesian networks can represent all direct *and indirect* dependencies of the problem domain by only explicitly showing the *direct* dependencies.

The following simple example, adapted from Taroni *et al.* 2006, p. 39, may illustrate the concept. The subject holds some prior beliefs about the fairness of a coin, which can be fair (heads and tails on opposite sides), tails only or heads only. H is the variable that represents this prior belief, and H = fair, H = only heads and H = only tails are the three mutually exclusive states it can take. The subject now observes the outcome of a first throw of this coin (she cannot examine the coin). The variable E_1 represents this evidence (while E_2, E_3 stand for the second and third toss), and it can take the states E_1 = head and E_1 = tail. This obviously tells the subject something about the fairness of the coin, and this in turn will influence her expectation that the next toss of the coin lands on heads (see Figure 1 a)). However, *given* that the coin is in any of its states, the outcome of the first toss will not tell the subject anything about the further outcomes. The variables E_1, E_2 and E_3 are *conditionally independent* given E. This knowledge of the conditional independencies is brought to the table by the human expert who knows that the first toss of a fair coin tells her nothing about the probable outcome of the second toss and allows a more parsimonious representation of the problem as given in Figure 1 b). Figure 1 b) also shows the (conditional) probability tables associated with each node of the network.

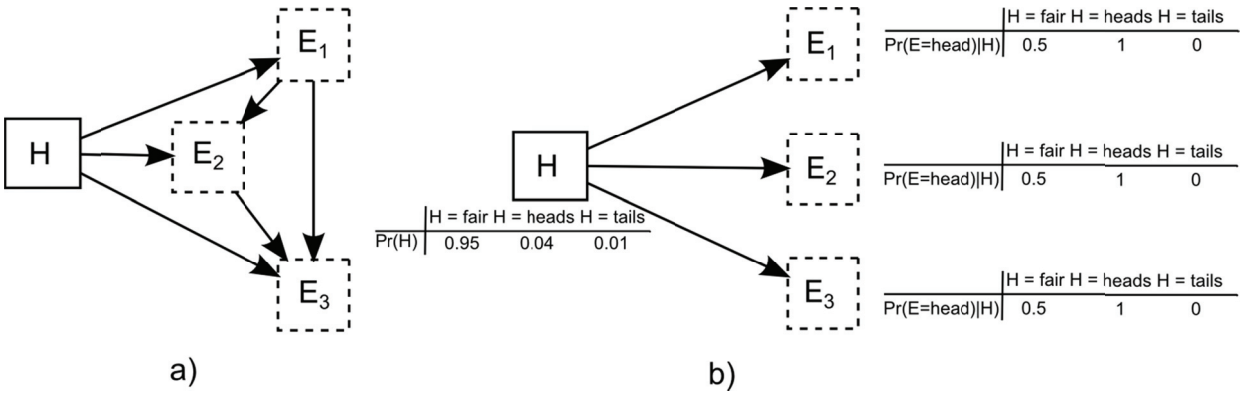

FIGURE 1: Bayesian network with all dependencies (Fig. 1a) and only the direct dependencies (Fig. 1b) represented by directed arcs

A Bayesian network is a correct representation of the problem domain if all variables A_1, …, A_{n-1} that have a *direct* influence on A_n are parents of A_n. If this is the case and the state of the parents of A_n is known, the variable A_n is independent of all its other ancestors. This means that in a Bayesian network,

$$\Pr(A_n|A_1,\dots,A_{n-1}) = \Pr\big(A_n\big|\text{parents}(A_n)\big) \tag{6}$$

holds (for a proof, see Charniak 1991, p. 55 seq.). This allows simplifying the chain rule, which in turn allows the calculation of the full joint probability distribution for the problem domain, to the following:

$$\Pr(A_1,\dots,A_n) = \prod_{i=1}^{n} \Pr(A_i \mid A_1,\dots,A_{i-1}) = \prod_{i=1}^{n} \Pr(A_i \mid \text{parents}(A_i)) \tag{7}$$

In the coin-tossing example, the full joint probability distribution can therefore be represented by

$$\Pr(H, E_1, E_2, E_3) = \Pr(H)\Pr(E_1|H)\Pr(E_2|H)\ \Pr(E_3|H)$$

which can be used to reconstruct the full joint probability distribution. This may not seem like a large simplification. However, the specification of the full joint probability distribution for the case reported below with 11 binary variables requires $2^{11} - 1 = 2047$ values, while the Bayesian network of the same case allows the reconstruction of the full joint probability distribution using just 22 (conditional) probabilities.

If the subject in the coin-tossing example wishes to condition her belief in H on the evidence E_1, E_2, E_3, Bayes' rule tells her to calculate

$$\Pr(H|E_1,E_2,E_3) = \frac{\Pr(H,E_1,E_2,E_3)}{\Pr(E_1,E_2,E_3)} = \frac{\Pr(H)\Pr(E_1|H)\Pr(E_2|H)\ \Pr(E_3|H)}{\sum_H(H,E_1,E_2,E_3)}.$$

For demonstration purposes, the actual calculation is carried out for a very simple example, i.e., for the case where the subject observes that all three tosses of the coin land on head. She must update her prior belief in the fairness of the coin, $\Pr(h_{fair}) = 0.95$, $\Pr(h_{heads}) = 0.04$, $\Pr(h_{tails}) = 0.01$, the following way (h = head)

$$\begin{aligned}\Pr(h_{fair}|e_{1\,h}, e_{2\,h}, e_{3\,h}) &= \frac{\Pr(h_{fair}, e_{1\,h}, e_{2\,h}, e_{3\,h})}{\Pr(e_{1\,h}, e_{1\,h}, e_{3\,h})} \\ &= \frac{\Pr(h_{fair})\ \Pr(e_{1\,h}|h_{fair})\Pr(e_{2\,h}|h_{fair})\Pr(e_{3\,h}|h_{fair})}{\sum_H(e_{1\,h}, e_{1\,h}, e_{3\,h})} \\ &= \frac{0.95 \times 0.5 \times 0.5 \times 0.5}{(0.95 \times 0.5 \times 0.5 \times 0.5) + (0.4 \times 1) + (0.01 \times 0)} \\ &= 0.748.\end{aligned}$$

That is, after observing three tosses that fall on heads in a row, her belief that the coin is fair is reduced from 0.95 to 0.75. For more complex queries, the calculation is tedious using paper and pencil even for small networks and impossible for large networks. Algorithms have been developed that perform these calculations efficiently for large networks (Pearl 1986; Lau-

ritzen & Spiegelhalter 1988). For certain classes of networks, an exact solution is impossible, but algorithms for approximate solutions exist (Darwiche 2009, p. 340 seq.).

For the user of Bayesian networks, knowledge of the algorithms is just as unnecessary as knowledge of the internal workings of a calculator is unnecessary for the use of a calculator (Fenton & Neil 2011, p. 131). It is sufficient to know that the algorithms have been accepted by the scientific community as correct, and that different implementations lead to the same results. There are a number of both commercial and free software programs available for probabilistic inference using Bayesian networks. All calculations for this article were performed with SamIam (Sensitivity Analysis, Modeling, Inference And More) 3.0, which is developed by the Automated Reasoning Group of Professor Adnan Darwiche at UCLA. This software is free and well documented by a number of scientific papers and a book (Darwiche 2009); however, the same results could have been obtained by any number of programs.

It must be noted that subjective probability theory is by no means the only "atomistic" model of evidence evaluation. Cohen's "inductive probabilities" (Cohen 1977), the evidentiary value model of Ekelöf/Halldén/Edman (Ekelöf 1964; Edman 1973; Halldén 1973) and Shafer-Dempster belief functions (Shafer 1976) are also "atomistic" models of the evaluation of evidence, in the sense that they require the assessment of the probative value of each item of evidence, and the overall assessment of the case is generated computationally by an algorithm. I do not wish to distract from the merits of these models. However, neither of these models allows the computation of a normative *degree of belief*; they do not purport to be models of belief formation, but rather models of *evidentiary support*. Therefore, their results are not directly comparable to a holistic degree of belief in the guilt of the defendant. Only subjective probability theory allows a meaningful comparison of a degree of belief that has been arrived at intuitively with one based on normative rules.

IV. Hypotheses

I sought to examine two main hypotheses. The first stems from the (empirically corroborated) prediction of PCS models of cognitive coherence that a holistic evaluation of evidence leads to inflated confidence in the truth of whichever hypothesis the subject accepts. I hypothesize that forcing the subject to assess the likelihoods for each individual item of evidence and integrating the obtained values using a Bayesian network would lead to reduced "coherence shifts". This is based on the observation that counterfactual thinking helps reduce coherence shifts (Simon 2004, p. 544). Thinking in a likelihood framework forces counterfactual thinking upon the subjects by making them consider that the observation may also have been made if the hypothesis to be tested was not true.

The second main hypothesis is that computing the posterior probability of guilt reduces the variability in the assessment of guilt compared to a holistic, intuitive assessment. This hypothesis is based on results by Schum & Martin who report that when the evaluation of evi-

dence is decomposed into individual items, inter-individual differences in the evaluation of the evidence are reduced (Schum & Martin 1982).

V. Experiment

A. Method

a. Participants

I invited 120 subjects to the Hermann Ebbinghaus lab at the University of Erfurt, Germany, for completion of a computer-based questionnaire. Sessions lasted about one hour. Subjects were paid € 6 for participation. Six subjects could not complete the questionnaire because of a computer malfunction. 16 subjects provided values that resulted in networks that could not be queried (see below, B. Results, for an explanation) and were excluded from further analysis, which leaves 98 subjects. The subjects were students between the ages of 19 and 47 with an average age of just under 24 (median 23). 72% were women. An overwhelming majority majors in pedagogy or psychology.

b. Material and Procedure

Subjects first read the instructions for completing the questionnaire and answered corresponding test questions before reading the case material. They were instructed that they could imagine a subjective probability of x% as the expectation of blindly drawing a red ball from an urn containing 100 balls, thereof x red balls. Additionally, the meaning of conditional probability was explained, and some examples were given that were not from the domain of legal evidence. It was explained that likelihoods need not add up to one.[1] Subjects were asked to state probabilities as percentages from 0% to 100%, this being more natural than the mathematical convention of bounding probabilities between 0 and 1.

Subjects then read the scenario of a case involving the (alleged) theft of money from a safe (the "Jason Wells/Hans H. case"), a scenario that has been used in a number of psychological studies (Simon *et al.* 2004; Glöckner & Engel 2008; Engel & Glöckner 2012; a full transcript has been published in the last two references). The scenario, of just over 700 words, describes Hans, a 34 year old married man with two children, employed at a construction company. Hans has recently been denied a promotion. He has a prior criminal record for attempted burglary at age 18, but has not since come into conflict with the law. One day, € 5,200 is missing from the company's safe. 8 people, among them Hans, have access to the safe, which was last opened at 7.14 pm. A technician testifies that he saw Hans leave the office in which the safe is located at about 7.15 pm. A surveillance video shows a car of the rare kind Hans drives leaving from the office building at 7.17 pm, but the license plate is illegible. Another witness, Silvia, testifies that she saw Hans at a school function at 8 pm wearing different clothes than the

1 This instruction was added based on the observation from a pre-test, in which a substantial number of subjects gave responses to the likelihood questions that always summed to 100%, which seems to imply that they (wrongly) thought that this must be the case.

ones he wore at work, and it would be difficult to get from the office to the school in less than 40 min at that time of day. The day after the disappearance of the money from the safe, Hans repays a bank loan of € 4,870. Hans claims he received this money from his sister-in-law who owns a flower shop, but he cannot produce a receipt for the transaction. He explains this by the practice in the flower business of "occasionally" doing business without issuing receipts.

The Hans case contains contradictory as well as missing evidence. Missing is the receipt, for which Hans offers an explanation, but Hans also fails to call his sister-in-law as a witness. The case is silent on why Hans does not offer the testimony of his sister-in-law, and the subjects are not expressly alerted to the omission.

After reading the scenario, subjects indicate whether they consider Hans guilty of taking the money or not. They then state their subjective probability of guilt ("holistic before"). Then, subjects indicate the subjective probability of guilt they think is required for a criminal conviction ("own standard") and after reading the definition of the criminal standard of proof used by the German Federal Supreme Court they indicate the subjective probability they think this standard requires ("legal standard"). They are then asked to give their prior probability of guilt. First, they are asked to state their prior belief given that Hans is one of eight people who have access to the safe ("objective prior" – this prior is of course also a subjective probability, but unlike the other subjective probabilities in this case, it is based on a known relative frequency). The subjects are then asked to state their prior belief for guilt given that Hans is one of eight people with access, has been denied a promotion and has a prior criminal record ("subjective prior" taking into account Hans' character and motive). Subjects also indicate their prior belief in Hans having received the money from his sister-in-law (the other root node of the network).

The likelihood ratio for each item of evidence was then elicited from the subjects using natural language questions. For example, for the witness statement of the technician, subjects have to answer the questions "How likely is it that the technician testifies he saw Hans leaving the office, given that Hans left the office?" and "How likely is it that the technician testifies he saw Hans leaving the office, given that Hans *did not* leave the office?". Subjects assessed a total of 11 likelihood ratios; this allowed the computation of two different versions of the Bayesian network (see below). At the end of the questionnaire, subjects were asked again what their holistic subjective probability for Hans' guilt was ("holistic after").

c. Computation of the Posterior Probability of Guilt

The posterior probability of guilt was computed for each subject using the parameters obtained from that subject and the structure of the Bayesian network given in Figure 2. Evidence variables, i.e., variables the state of which is observed, are shown in dashed rectangles. The hypothesis variable (or "query variable", Darwiche 2009, p. 84) is the variable of interest; it is shown in a rectangle with a thick border. Intermediate variables are variables that cannot be observed and mediate the influence of the evidence variables on the hypothesis variable; they

are shown in solid line rectangles. The two evidence variables "refused promotion" and "old criminal record" are not required for the computation of the network, as their state is known and they are parents of the hypothesis variable. They are included in Figure 2 for sake of completeness. The evidence regarding motive and character is of course reflected in the subjective prior probability for Hans' guilt and does therefore influence the computed posterior probability of guilt, but it needs not be added as distinct variables to the network.

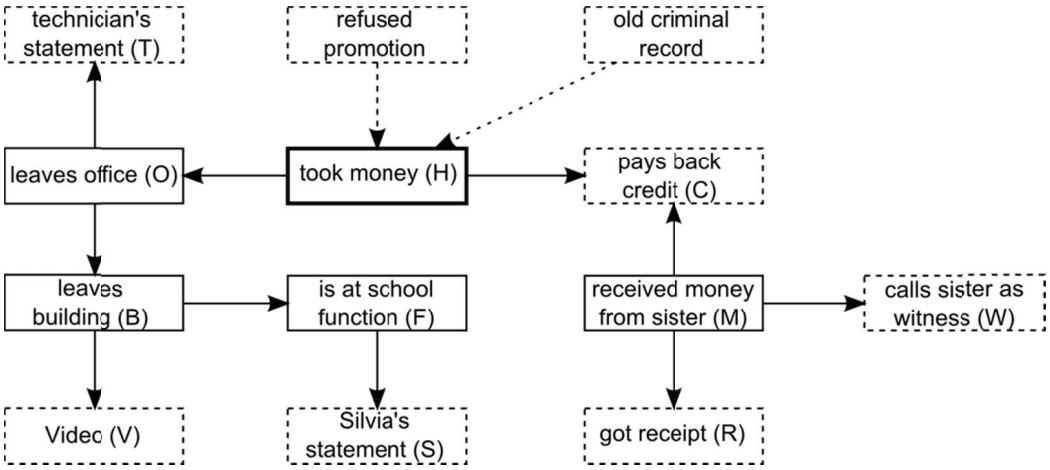

Figure 2: Bayesian network for the Jason Wells/Hans H. case. Evidence variables with dashed borders, intermediate variables with solid borders, hypothesis variable with thick border

Given the structure of the network in Figure 2, the full joint probability distribution for the network can be factorized as follows:

$$\Pr(H, M, O, B, F, T, V, S, R, W, C)$$
$$= \Pr(H)\,\Pr(M)\,\Pr(O|H)\,\Pr(B|O)\,\Pr(F|B)\,\Pr(T|O)\,\Pr(V|B)\,\Pr(S|F)\,\Pr(R|M)\,\Pr(W|M)\,\Pr(C|H, M)$$

The states of the evidence variables T, V, S and C are observed as true, and the states of the evidence variables R and W are observed as false. Known states are indicated with lower-case letters with the indices t = true and f = false. The query of interest is therefore

$$\Pr(h_t|M, O, B, F, t_t, v_t, s_t, r_f, w_f, c_t)$$

$$= \frac{\Pr(h_t)\,\Pr(M)\,\Pr(O|h_t)\,\Pr(B|O)\,\Pr(F|B)\,\Pr(t_t|O)\,\Pr(v_t|B)\,\Pr(s_t|F)\,\Pr(r_f|M)\,\Pr(w_f|M)\,\Pr(c_t|h_t, M)}{\sum_H \sum_M \sum_O \sum_B \sum_F \Pr(t_t, v_t, s_t, r_f, w_f, c_t)}$$

It would be highly impractical to carry out the actual summations required for the solution by hand, and all calculations were therefore performed using the software SamIam 3.0. A total of four posterior subjective probabilities of $\Pr(h_t)$ were computed for each subject: the first one using the structure given above with all the evidence variables instantiated ("computed posterior 1") and the second one with the same structure, but without the evidence variable W (calls sister-in-law as witness) being instantiated ("computed posterior 2"). The second and third posterior probabilities were computed with an alternative structure of the network in which the variable S (school function) is a child of H (took money). This is incorrect, as the

probability of reaching the school in time is not *directly* dependent on taking the money, only on leaving the building in time (simply put, Hans is not slower across town with money in his pockets), but it is an intuitive representation of the variables' dependency structure. In this version of the network, Pr(F|H) replaces Pr(F|B), everything else remaining the same. Again, two posterior probabilities Pr(h_t) were computed, one with the evidence variable W instantiated ("alternative computed posterior 1"), one without instantiation of W ("alternative computed posterior 2").

B. Results

Of the 114 subjects who completed the questionnaire, 16 gave values for the likelihoods that made computation of the posterior for Pr(h_t) impossible. This occurs when subjects indicate probabilities that are inconsistent with the evidence. For example, one subject indicated that the probability that Hans would be at the school function *whether or not* he left the office building at 7.17 pm was 0%. She also indicated that the probability of Silvia testifying that Hans was at the school function, given that Hans was *not* at the school function, was 0%. Under these assumptions, it is *impossible* that Silvia testifies that Hans is at the school function, but we *know* that Silvia testified to this. Therefore, the conditional probabilities are inconsistent with the evidence and the network cannot be queried. The 16 subjects with networks that could not be queried were excluded from further analysis. Of those who were excluded, 11 (69%, versus 62% of the non-excluded subjects) would have convicted Hans. The average holistic probability of guilt for those 11 subjects was 80.2%, which is not significantly different from the 80.4% average holistic probability of guilt for the non-excluded convictors. The average holistic probability of guilt for those 5 excluded subjects who acquitted Hans was 23.1%, which is below the 45% for the non-excluded acquitters. If anything, including these 16 subjects in the analysis would therefore have increased the observed coherence shift.

61 subjects (62%) found Hans guilty of taking the money ("convictors") and 37 acquitted him ("acquitters"). Table 1 reports the average values of the objective prior probability, the subjective prior probability, the holistic probability of guilt given before and after answering the likelihood questions, the computed posterior of guilt using the first version of network with all the evidence variables instantiated, the computed posterior using the first version of the network without the variable W (calls sister-in-law as witness), the computed posterior using the alternative version of the network with all evidence variables instantiated and the computer posterior using the alternative version of the network without instantiation of W. The holistic posteriors of those who convict elicited before the likelihood questions ("holistic before") are significantly different from all the computed posteriors (all *ps*<0.05 using a two tailed paired t-test), while those elicited after the likelihood questions ("holistic after") are significantly different only from the computed posterior 2 and the alternative computed posterior 2. Either of the holistic posteriors of those who acquit is reliably different from the computed posterior 1 and the alternative computed posterior 1 (*ps*<0.05 using a two tailed paired t-test),

but not from the computed posterior 2 and the alternative computed posterior 2 (both *ps*>0.11).

Table 1: Mean prior, holistic and computed posterior probabilities, by subjects who convict and acquit (standard deviation)

	Objective prior	Subjective prior	Holistic before	Holistic after	Computed post. 1	Computed post. 2	Alt. computed post. 1	Alt. computed post. 2
Convictors	12.8 (2.5)	24.2 (18.4)	80.4 (20.6)	75.7 (19.6)	69.0 (36.3)	50.5 (33.3)	67.9 (37.1)	46.7 (33.9)
Acquitters	12.6 (2.4)	20.6 (18.3)	45.0 (27.0)	45.1 (26.6)	62.8 (36.8)	40.5 (29.7)	59.7 (37.2)	34.5 (28.8)
Average	12.8 (2.5)	22.9 (16.7)	67.0 (28.8)	64.2 (26.8)	66.7 (36.4)	46.7 (32.2)	64.8 (37.2)	42.1 (32.4)
Difference	0.2	3.6	35.4***	30.4***	6.2	10.0+	8.2	12.2+

*** p<0.001, + p<0.1 (using a two sided t-test).[2]

Taking into account or ignoring that Hans failed to call his sister-in-law results in significant differences in the mean computed posteriors. Within a random effect regression with the computed posterior as the dependent variable the dummy variable for the instantiation of variable W of the network has predictive power (*b* = 19.98, z(98)=6.92, p <.001). No effect for the verdict (convict or acquit) was found (*b* = -8.1, z(98)=1.2, p =.214).

Figure 3 graphically displays the data from Table 1. It shows how both convictors and acquitters share almost the same priors, but differ strongly in their holistic posterior probability of guilt. Answering the likelihood questions only marginally decreases the holistic posterior for the convictors and has no effect on the holistic posterior of the acquitters. However, computing the posterior using the likelihoods given by the subjects greatly *increases* the posterior for the acquitters and *decreases* the posterior for the convictors, bringing the two groups closely together. Computing the posterior without taking into account that Hans did not call his sister-in-law as a witness further decreases the posterior probability for both groups. However, the main effect, the closing of the gap between the assessments of the case, remains.

2 To check for robustness, a Wilcoxon rank-sum test was also used. The results remain the same.

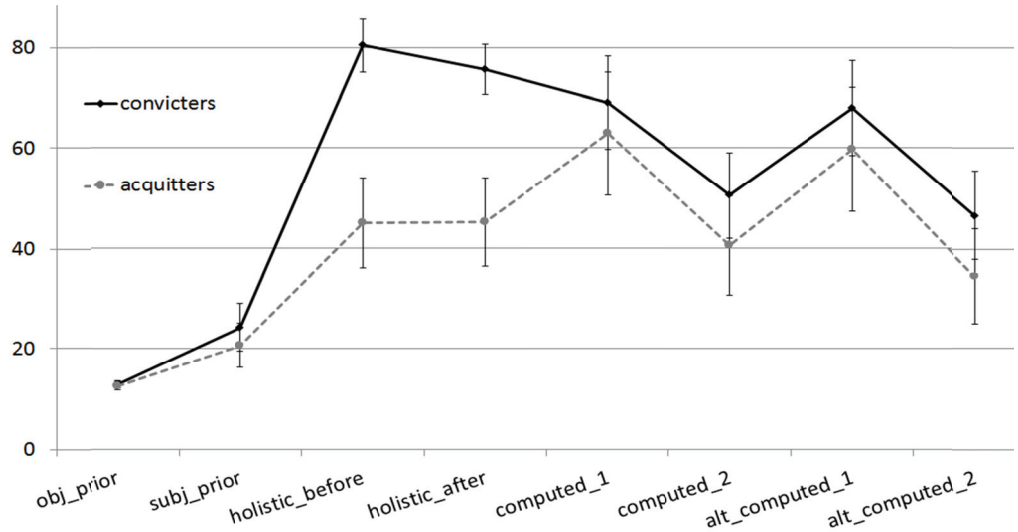

Figure 3: Mean posterior probabilities, by subjects who convict and acquit (error bars indicate 95% confidence intervals)

Table 2 shows the average likelihoods for convictors and acquitters for each item of evidence and each intermediate hypothesis. With the exception of the difference in means for the likelihood $Pr(o_t|h_t)$ that Hans leaves the office at 7.15 pm, given that he took the money, and the likelihood $Pr(t_t|o_f)$ that the technician testifies that he saw Hans leaving the office, given that Hans did not leave the office, which were significant at a level of $p<0.05$ and $p<0.1$ respectively, none of the differences in mean likelihoods are statistically significant.

Table 2: Mean likelihoods, by subjects who convict and acquit (standard deviation)

	$Pr(o_t\|h_t)$	$Pr(o_t\|h_f)$	$Pr(t_t\|o_t)$	$Pr(t_t\|o_f)$	$Pr(b_t\|o_t)$	$Pr(b_t\|o_f)$	$Pr(v_t\|b_t)$	$Pr(v_t\|b_f)$
Convictors	74.7	41.8	89.3	23.6	78.6	29.7	87.4	22.8
	(25.5)	(28.8)	(15.7)	(25.5)	(22.1)	(26.4)	(24.7)	(32.8)
Acquitters	63.7	40.5	85.5	33.2	77.1	36.6	90.6	34.5
	(28.2)	(28.4)	(19.8)	(29.6)	(22.9)	(28.3)	(16.9)	(39.4)
Difference	11.0**	1.3	3.8	9.6+	1.5	6.9	3.2	11.7

** p<0.05, + p<0.1 (using a two sided t-test).[3]

Table 2 (cont'd): Mean likelihoods, by subjects who convict and acquit (standard deviation)

	$Pr(f_t\|b_t)$	$Pr(f_t\|b_f)$	$Pr(s_t\|f_t)$	$Pr(s_t\|f_f)$	$Pr(r_t\|m_t)$	$Pr(r_t\|m_f)$	$Pr(w_t\|m_t)$	$Pr(w_t\|m_f)$
Convictors	53.9	61.7	93.4	19.2	35.8	7.4	90.6	35.0
	(29.6)	(31.8)	(14.7)	(26.2)	(31.0)	(15.2)	(19.4)	(29.6)
Acquitters	47.6	65.9	87.2	27.1	34.5	9.4	86.6	35.1
	(28,3)	(28.3)	(20.1)	(26.9)	(28.9)	(17.4)	(22.9)	(29.1)
Difference	6.3	4.2	6.3	7.9+	1.3	2.0	4.0	0.1

3 To check for robustness, a Wilcoxon rank-sum test was also used. The results remain the same.

Table 2 (cont'd): Mean likelihoods, by subjects who convict and acquit (standard deviation)

	$Pr(f_t\|h_t)$	$Pr(f_t\|h_f)$	$Pr(c_t\|h_t, m_t)$	$Pr(c_t\|h_f, m_t)$	$Pr(c_t\|h_t, m_f)$	$Pr(c_t\|h_f, m_f)$
Convictors	53.1 (29.6)	76.9 (24.3)	62.5 (35.8)	66.1 (30.4)	63.6 (25.9)	14.4 (25.8)
Acquitters	45.9 (27.4)	75.4 (24.5)	67.8 (36.5)	74.6 (27.1)	64.1 (31.1)	18.6 (29.0)
Difference	7.2	1.5	5.3	8.5	0.5	4.2

Figure 4 shows the likelihood *ratios* for all items of evidence and all intermediate hypotheses, which were calculated using the averages for the likelihoods from Table 2. For example, the likelihood ratio "ratio_office" for the convictors was computed by dividing the average likelihood $Pr(o_t|h_t) = 74.7\%$ by the average likelihood $Pr(o_t|h_f) = 41.8\%$ (all values from Table 2), which results in a likelihood ratio of 1.785. In other words, it is believed to be 1.78 times as likely that Hans leaves the office at exactly 7.15 pm if he took the money as if he did not take the money. For the variable C (pays back credit), which has two parents H (took money) and M (received money from sister-in-law), two likelihood ratios were computed: $Pr(c_{true}|h_{true}, m_{true})/Pr(c_{true}|h_{false}, m_{true})$ ("ratio_credit 1") and $Pr(c_{true}|h_{true}, m_{false})/Pr(c_{true}|h_{false}, m_{false})$ ("ratio_credit 2").

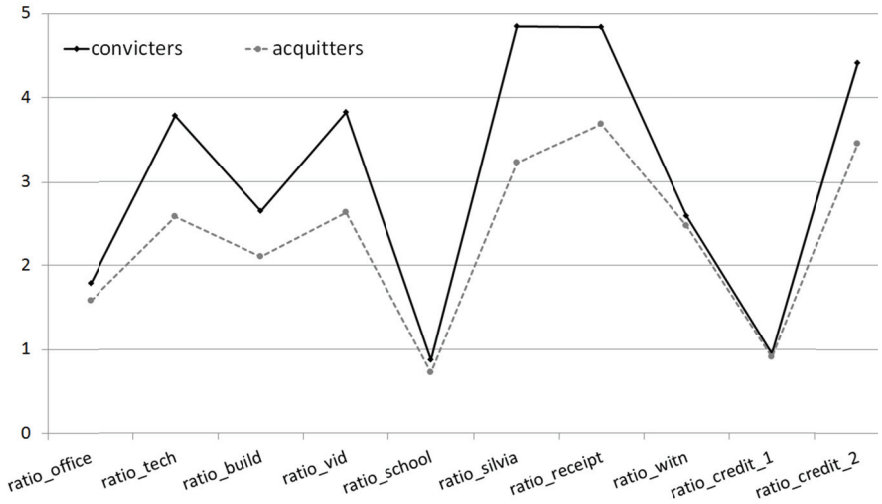

Figure 4: Mean likelihood ratios for each item of evidence and each intermediate hypothesis, by subjects who convict and acquit

Table 3 shows the subjective posterior probability of guilt that the subjects believe is required for a conviction in a criminal case ("own standard") and the subjects' interpretation of the criminal standard of proof in Germany as a threshold subjective probability ("legal standard"). The subjects read the following definition of the standard of proof in criminal matters commonly used by the German Federal Supreme Court (e.g. BGH, 30 July 2009 – 3 StR 273/09 = BeckRS 2009, 25658; translation into English by the author):

> "The conviction of the judge does not require an absolute certainty that excludes other possibilities with logical necessity. An adequate degree of certainty that overcomes reasonable doubt is sufficient. The judge is not prohibited from drawing possible, albeit not cogent inferences, from facts if such inferences are supported."

Table 3: Mean subjective probability thresholds required for a conviction, by subjects who convict and acquit (standard deviation)

	own standard	own standard (w/o 100%)	legal standard
Convictors	93.1 (15.7)	87.6 (19.5)	73.1 (13.0)
Acquitters	95.9 (6.2)	93 (6.8)	76.4 (20.4)
Difference	2.8	5.4	3.3

To investigate whether the high unguided standard of proof was caused by the 42 of the subjects who indicated that a certainty of 100% is necessary, I excluded these subjects for an additional calculation of the mean own standard, because under no legal rule is absolute certainty a requirement for a conviction. Excluding these subjects leads to the average threshold posteriors reported in the middle column of Table 3. There are no significant differences in the mean thresholds required for conviction between those who acquit and those who convict.

Table 4 compares the posteriors (column headings) with the probability thresholds required for a conviction (line headings) and counts instances where the posterior meets or exceeds the threshold. In the first two lines of Table 4, the threshold probability stated by each subject is compared with his or her individual (holistic and computed) posterior. In columns one and two, the comparison standard is the subject's own holistic posterior belief in guilt, expressed before answering the likelihood question. In columns three and four, the comparison standard is the subject's own computed posterior of guilt taking into account the adverse inference based on the missing witness. In columns five and six, the comparison standard is the subject's own computed posterior of guilt without the adverse inference based on the missing witness.

If the posterior exceeds the personal threshold value, the subject should (according to his or her own standard) convict or, if not, acquit. The top left cell of Table 4 shows that the holistic posterior (before answering the likelihood questions) for only 11 subjects exceeds their own personal threshold value for a conviction. 42 subjects (38 who actually convicted and 4 who acquitted) have holistic posteriors that meet or exceed their own interpretation of the legal standard of proof. The last two lines of Table 4 compare the individual posterior probabilities of guilt with the average threshold probabilities and the average legal standard for all subjects and count all instances where the posterior meets or exceeds the average threshold.

Table 4: Instances of posteriors meeting or exceeding the threshold probability for conviction

	Holistic posterior (bf.)		Comp. post. 1 (w/ adv. inf.)		Comp. post. 2 (w/o adv. inf.)	
	Convictors (n=61)	Acquitters (n=37)	Convictors (n=62)	Acquitters (n=37)	Convictors (n=62)	Acquitters (n=37)
Indiv. own std.	11	0	19	8	10	1
Indiv. legal std.	38	4	36	16	19	7
Avg. own std.	19	2	30	14	11	3
Avg. legal std.	46	6	39	19	20	5

Figure 5 shows the instances where different posteriors meet or exceed different threshold levels for conviction as *proportions* of the total number of possible convictions. 62% of the subjects actually convicted Hans. As is evident, only a fraction of those 61 subjects *should have* convicted him had they adhered to their own standard of proof.

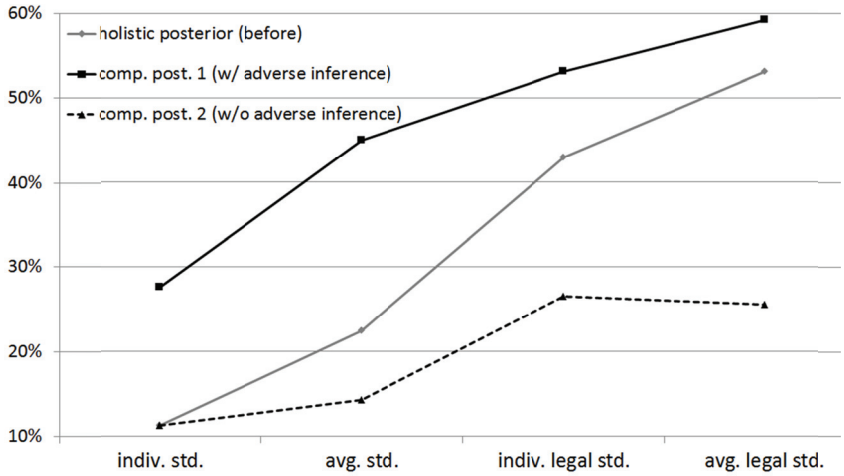

Figure 5: Proportion of cases where the posterior meets or exceeds the threshold probability required for a conviction

C. Discussion

The holistic posterior probability for the defendant's guilt by the subjects convicting the defendant is roughly twice as high as the holistic posterior probability of guilt for the subjects acquitting the defendant. This closely replicates results by Glöckner & Engel 2008 and indicates that under a holistic evaluation of the evidence, convictors and acquitters really "interpreted the same case in completely different ways" (Glöckner & Engel 2008, p. 13), as predicted by PCS models of cognitive coherence. However, when the subjects atomistically evaluate the evidence, this is no longer the case: computing the posterior probability of guilt using a Bayesian network with the parameters for the prior probabilities and the likelihoods

obtained from the subjects makes the difference in the evaluation of the case between the convictors and acquitters largely disappear. Merely answering the likelihood questions is not sufficient to achieve this effect; it only marginally decreases the posterior probability of guilt for the convictors and has no effect on the posterior probability of guilt of the acquitters. The data therefore support the first main hypothesis: an atomistic evaluation of evidence in a likelihood framework leads to the disappearance of coherence shifts.

This conclusion is further supported by the data for the likelihood ratios for each item of evidence. The differences of the mean likelihoods for convictors and acquitters are not significant at the $p<0.05$ level for all except one likelihood. A comparison of the likelihood ratios "ratio credit 1" and "ratio credit 2" shows that the subjects correctly interpret the repayment of the loan – the day after the disappearance of the money from the safe – as incriminating evidence, given that Hans has *not* received the money from his sister-in-law ("ratio credit 2"); however, the subjects assign no probative value to the repayment given that Hans *has* received the money from his sister ("ratio credit 1"). In the latter case, the repayment is adequately explained even without Hans having taken the money from the safe.

It is also noteworthy that on average, all the evidence was judged to be of limited strength. None of the likelihood ratios computed by dividing the average likelihoods reached more than 5. According to the verbal scale for forensic evidence suggested by Evett et al., a likelihood ratio of 1 to 10 can be verbally expressed as "limited evidence to support" (Evett *et al.* 2000, 236). A different picture emerges when individual likelihood ratios are computed for each subject, using the likelihoods provided by that subject. These likelihood ratios were sometimes very large, indicating very strong evidence (see below, D. Limitations, for a discussion of the large observed inter-individual differences in the assessment of the likelihoods).

Taking into account or ignoring that Hans failed to call his sister-in-law as a witness results in a large difference of about 20 percentage points in the computed posteriors. It reflects the intuition that not calling the witness allows inferring that the witness' testimony would be unfavourable for Hans. This intuition is reflected in US case law going back to *Graves vs. United States*, where the US Supreme Court stated: "The rule even in criminal cases is that if a party has it peculiarly within his power to produce witnesses whose testimony would elucidate the transaction, the fact that he does not do it creates the presumption that the testimony, if produced, would be unfavorable." (*Graves vs. United States*, 150 U.S. 118, 121 [1893]). A party has the power to produce a witness if "[it] had the physical ability to locate and produce the witness and there was such a relationship, in legal status or on the facts as claimed by the party as to make it natural to expect the party to have called the witness" (*Thomas v. United States*, 447 A.2d 52, 57 [D.C. 1982]). Given that the witness in question is Hans' sister-in-law and would have first-hand knowledge of the relevant issue whether Hans received the money from her, these conditions appear to be met. The data supports the conclusion that the subjects took the mere omission of calling the sister-in-law as a witness as evidence against the truth of the proposition that Hans received the money to pay back the credit from his sister-in-law, which in turn increases the probability of Hans having taken the money because the alterna-

tive explanation becomes less probable. The fact that Bayesian networks can model such relatively complex chains of inference is one of their strengths.

However, the adverse inference is not permissible in this case, since the prosecutor also could have called the witness. US courts have applied the missing witness inference rule in criminal cases, provided that the state cannot reasonably locate the missing witness (*U.S. v. Anchondo-Sandoval*, 910 F.2d 1234, 1238 [5th Cir. 1990]). This prerequisite is most probably (the scenario is silent on the issue) not met because there are no reasons to think that the state could not have located Hans' sister-in-law and called her as a witness. In defense of the subjects it must be stressed that they were only asked about the likelihood of *Hans* not calling the witness given that he received/did not receive the money from his sister-in-law. A fairer question would have been how likely it was whether Hans *or the prosecution* did not call the witness given that Hans did not receive the money from his sister-in-law.

There are no significant differences in the threshold probability required for a conviction between those who convicted and those who acquitted the defendant. Interestingly, the mean unguided estimate of the required threshold level for a conviction in criminal matters is closer to the values of well above 90% that are stated in the German legal literature (e.g. Hoyer 1993, p. 439) than the subjects' interpretation of the threshold level required by the German Federal Supreme Court. However, the subjects are inconsistent with their own standards: 50 subjects convicted Hans although their stated holistic posterior probability of guilt did not exceed their own stated threshold probability for a conviction, and still 23 convicted although their holistic posterior did not even meet their own understanding of the legal standard of proof (see Table 3). Arguably, the comparison with the *average* legal standard is most appropriate, as the legal standard of proof should not vary between decision makers. Comparing the computed posteriors with the average legal standard shows that drawing an adverse inference from the failure to call the witness leads, as expected, to a substantially larger proportion of convictions.

The average holistic posterior for the guilt of the defendant of those who convict (80.4%) is above the average legal standard required for a conviction as expressed by the subjects (74.4%). However, the average computed posterior probability of guilt *even for the convictors* just barely exceeds 50% (50.5%) if one does not draw an adverse inference from the missing witness, as would be correct in this case. Compared to the average of the legal standard in criminal matters in Germany as expressed by the subjects, Hans should not have been convicted based on the item-by-item assessment of the evidence. A posterior probability of guilt of 50.5% also fails to exceed any reasonable quantification of the "beyond reasonable doubt" standard of proof in criminal matters of US law. While there is considerable inter-individual variability in the expression of the "beyond reasonable doubt" standard as a degree of probability (see Hastie 1993, p. 101 seq.) it is generally understood to require a much higher probability than the "preponderance of the evidence" standard of just above 50% used in civil cases (Lillquist 2002, p. 94). Whether a quantification is desirable at all is the subject of an ongoing debate; while scholars have long advocated the use of a numerical definition of the

standard of proof, courts have remained hostile to attempts at quantification (see Tillers & Gottfried 2007).

The second main hypothesis is not supported by the data: based on Schum & Martin 1982, it was hypothesized that an item-by-item assessment of the evidence would lead to a reduction in the variance of the posterior probability of guilt. This was evidently not the case. As the standard deviations for the likelihoods (Table 2) indicate, there was actually higher variance in the assessment of the conditional probabilities than in the assessment of the holistic posterior probability of guilt. This is because many subjects chose extreme values of 0% or 100% for the likelihoods. These extreme values for the evidence and intermediate variables carry over into the computed posteriors, which also show higher standard deviations than both the holistic posteriors (see Table 1). It has long been thought that posterior probabilities computed using Bayesian networks are robust to changes in the values for the evidence and intermediary variables (Pradhan *et al.* 1996); however, this is not generally true. Networks with *extreme* values for the evidence and intermediate variables (i.e. values close to the bounds of 0 and 1) and *intermediate* values on the query variable(s) are sensitive to changes in the parameters of the evidence variables (Chan & Darwiche 2002). Intuitively, this can be explained by considering that a small absolute change in an extreme value of a likelihood, let's say from 0.001 to 0.01, increases the likelihood ratio by an order of magnitude, while the same small change in an intermediate probability, say from 0.601 to 0.61, has almost no influence on the likelihood ratio (assuming all else being equal).

I can only speculate as to why the results from Schum & Martin 1982 could not be replicated. A plausible explanation is that the 20 subjects of Schum & Martin gave a total of 16,000 probability assessments (800 per subject) over the course of several days and assessed the same evidence repeatedly (Schum & Martin 1982, p. 127 seq.). This may have induced learning and thereby higher consistency. The subjects in this study, on the other hand, were unfamiliar with the task of assigning numerical values to degrees of belief and had little opportunity for learning. This unfamiliarity with a task that is known to be difficult may have led to the great observed variance.

D. Limitations and further research

As should have become evident from the discussion, the main limitation of this study stems from the old maxim that averages can be deceiving. The average values for the parameters of the network and the average computed posteriors support the main hypothesis. Looking at the data for the individual subjects reveals great inter-individual differences. Since it is desired that the evaluation of evidence in a judicial context is predictable, these inter-individual differences should be reduced if Bayesian networks are to be a useful tool for the fact finder in cases where there are no relative frequencies that could inform the subjective probabilities. Further research should therefore explore whether more sophisticated elicitation techniques

for the (conditional) probabilities lead to less inter-individual variability (for an overview of different elicitation techniques, see O'Hagan *et al.* 2006).

The second limitation of this study is that the structure of the network was designed by the experimenter and therefore the same for all subjects. As the direct dependencies which structure the model are based on the expert's knowledge and assumptions about the workings of the world, different experts may structure the problem differently. Further research should explore whether the main effect, the large reduction in the difference in the posterior probability of guilt for the convictors and acquitters, remains if not only the parameters for the network, but also the network *structure* is elicited from the subjects.

VI. Conclusion

This study is the first to empirically demonstrate an advantage of using a Bayesian network for the evaluation of evidence in a case where there are no relative frequencies that could form the basis for assessing the probative value of the evidence. The study shows that the large difference between the posterior subjective probability of guilt of judges who convict and judges who acquit largely disappears when the posterior probability of guilt is computed using a Bayesian network parameterised with the values obtained from the judges. This result is important because the posterior degree of belief in the guilt of the defendant is the relevant standard of proof in most legal systems. Forcing coherence in the sense of subjective probability theory on the partial beliefs of the judge using a Bayesian network suppresses the polarization of evidence observed in the holistic evaluation of evidence and reduces the resulting inflated confidence in having made the right choice. It makes transparent that certainty is often unattainable in legal fact finding, and that the subjective feeling of certainty is mostly an illusion.

References

AITKEN C.G. (2012) An introduction to a debate. *Law, Probability and Risk,* first published online: 3 September, 2012.

AMAYA A. (2008) Justification, Coherence, and Epistemic Responsibility in Legal Fact-Finding. *Episteme* **5,** 306–319.

BROWNSTEIN A.L. (2003) Biased predecision processing. *Psychological Bulletin* **129,** 545–568.

CALLEN C.R. (1982) Notes on a grand illusion: some limits on the use of Bayesian theory in evidence law. *Indiana Law Journal* **57,** 1–44.

CARLSON K.A. & RUSSO J.E. (2001) Biased interpretation of evidence by mock jurors. *Journal of Experimental Psychology: Applied* **7,** 91–103.

CHAN H. & DARWICHE A. (2002) When do numbers really matter? *Journal of Artificial Intelligence Research* **17,** 265–287.

CHARNIAK E. (1991) Bayesian Networks without Tears. *AI Magazine* **12,** 50–61.

CHRISTENSEN D. (2007) Putting logic in its place. Formal constraints on rational belief. Clarendon Press, Oxford.

COHEN L.J. (1977) The probable and the provable. Clarendon Press, Oxford.

DARWICHE A. (2009) Modeling and reasoning with Bayesian networks. Cambridge University Press, Cambridge, New York.

EDMAN M. (1973) Adding independent pieces of evidence. In *Modality and Morality and Other Problems of Sense and Nonsense. Festskrift till Sören Halldén* (ed. B. Hansson), pp. 180–188. Gleerup, Lund.

EDWARDS W. (1991) Influence Diagrams, Bayesian Imperialism, and the Collins Case: An Appeal to Reason. *Cardozo Law Review* **13,** 1025–1074.

EKELÖF P.O. (1964) Free Evaluation of Evidence. *Scandinavian studies in law* **8,** 45–66.

ENGEL C. & GLÖCKNER A. (2012) Role-Induced Bias in Court: An Experimental Analysis. *Journal of Behavioral Decision Making,* first published online: 11 April 2012.

EVETT I.W., JACKSON G., LAMBERT J.A. & MCCROSSAN S. (2000) The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice* **40,** 233–239.

FENTON N. & NEIL M. (2011) Avoiding Probabilistic Reasoning Fallacies in Legal Practice using Bayesian Networks. *Australian Journal of Legal Philosohpy* **36,** 114–151.

FINETTI B. DE (1937) La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré* **7,** 1–68.

GLÖCKNER A., BETSCH T. & SCHINDLER N. (2010) Coherence shifts in probabilistic inference tasks. *Journal of Behavioral Decision Making* **23,** 439–462.

GLÖCKNER A. & ENGEL C. (2008) Can we trust intuitive jurors? An experimental analysis. *Preprints of the Max Planck Institute for Research on Collective Goods* **36**.

GOOD I.J. (1950) Probability and the weighing of evidence. Charles Griffin, London.

GOOD I.J. (1983) Some logic and history of hypothesis testing. In *Good thinking. The foundations of probability and its applications* (ed. I. J. Good), pp. 129–148. Univ. of Minnesota Press, Minneapolis.

HACKING I. (2008) An introduction to probability and inductive logic. Cambridge University Press, Cambridge.

HALLDÉN S. (1973) Indiciemekanismer. *Tidskrift for Rettsvitenskap* **86,** 55–64.

HASTIE R. (1993) Algebraic models of juror decision processes. In *Inside the Juror. The Psychology of Juror Decision Making* (ed. R. Hastie), pp. 84–115. Cambridge University Press, Cambridge.

HOLYOAK K.J. & SIMON D. (1999) Bidirectional reasoning in decision making by constraint satisfaction. *Journal of Experimental Psychology: General* **128,** 3–31.

HOPE L., MEMON A. & MCGEORGE P. (2004) Understanding Pretrial Publicity: Predecisional Distortion of Evidence by Mock Jurors. *Journal of Experimental Psychology: Applied* **10,** 111–119.

HOYER A. (1993) Der Konflikt zwischen richterlicher Beweiswürdigungsfreiheit und dem Prinzip "in dubio pro reo". *ZStW* **105,** 523–556.

JENSEN F.V. & NIELSEN T.D. (2007) Bayesian networks and decision graphs. Springer, New York.

JUCHLI P., BIEDERMANN A. & TARONI F. (2012) Graphical probabilistic analysis of the combination of items of evidence. *Law, Probability and Risk* **11,** 51–84.

KADANE J.B. & SCHUM D.A. (1996) A probabilistic analysis of the Sacco and Vanzetti evidence. Wiley, New York.

KAYE D.H. (1988) A first look at "second order evidence". In *Probability and inference in the law of evidence. The uses and limits of Bayesianism* (ed. P. Tillers), pp. 177–183. Reidel, Dordrecht.

LAGNADO D. (2011) Thinking about Evidence. In *Evidence, inference and enquiry* (eds. P. Dawid, W. L. Twining & M. Vasilaki), pp. 183–223. Oxford University Press, Oxford.

LAURITZEN S.L. & SPIEGELHALTER D.J. (1988) Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems. *Journal of the Royal Statistical Society. Series B (Methodological)* **50,** 157–224.

LILLQUIST E. (2002) Recasting Reasonable Doubt: Decision Theory and the Virtues of Variability. *U.C. Davis Law Review* **36,** 85–197.

LUNDBERG C.G. (2004) Modeling and predicting emerging inference-based decisions in complex and ambiguous legal settings. *European Journal of Operational Research* **153,** 417–432.

O'HAGAN A., BUCK C.E., DANESHKHAH A., EISER J.R., GARTHWAITE P.H., JENKINSON D.J., OAKLEY J.E. & RAKOW T. (2006) Uncertain judgements. Eliciting experts' probabilities. Wiley, Chichester.

PARK R.C., TILLERS P., MOSS F.C., RISINGER D.M., KAYE D.H., ALLEN R.J., GROSS S.R., HAY B.L., PARDO M.S. & KIRGIS P.F. (2010) Bayes Wars Redivivus. An Exchange. *International Commentary on Evidence* **8,** 1–38.

PEARL J. (1986) Fusion, Propagation, and Structuring in Belief Networks. *Artificial Intelligence* **29,** 241–288.

PEARL J. (1988) Probabilistic reasoning in intelligent systems. Networks of plausible inference. Morgan Kaufmann Publishers, San Francisco, Calif.

PENNINGTON N. & HASTIE R. (1992) Explaining the evidence. Tests of the Story Model for juror decision making. *Journal of Personality and Social Psychology* **62,** 189–206.

PRADHAN M., HENRION M., PROVAN G., DEL FAVERO B. & HUANG K. (1996) The sensitivity of belief networks to imprecise probabilities. An experimental investigation. *Artificial Intelligence* **85,** 363–397.

READ S.J., VANMAN E.J. & MILLER L.C. (1997) Connectionism, Parallel Constraint Satisfaction Processes, and Gestalt Principles. (Re) Introducing Cognitive Dynamics to Social Psychology. *Personality and Social Psychology Review* **1,** 26–53.

ROBERTSON B. & VIGNAUX G. (1992) Taking fact analysis seriously. *Michigan Law Review* **91,** 1442–1464.

RUSSO J.E., CARLSON K.A., MELOY M.G. & YONG K. (2008) The goal of consistency as a cause of information distortion. *Journal of Experimental Psychology: General* **137,** 456–470.

SCHUM D.A. & MARTIN A.W. (1982) Formal and Empirical Research on Cascaded Inference in Jurisprudence. *Law & Society Review* **17,** 105–151.

SHAFER G. (1976) A mathematical theory of evidence. Princeton Univ. Press, Princeton, N.J.

SIMON D. (2004) A Third View of the Black Box. Cognitive Coherence in Legal Decision Making. *The University of Chicago Law Review* **71,** 511–586.

SIMON D., SNOW C.J. & READ S.J. (2004) The Redux of Cognitive Consistency Theories. Evidence Judgments by Constraint Satisfaction. *Journal of Personality and Social Psychology* **86,** 814–837.

TARONI F., AITKEN C.G., GARBOLINO P. & BIEDERMANN A. (2006) Bayesian networks and probabilistic inference in forensic science. Wiley, Chichester.

THAGARD P. (1989) Explanatory coherence. *Behavioral and Brain Sciences* **12,** 435–502.

THAGARD P. (2000) Coherence in Thought and Action. MIT Press, Cambridge (Mass.).

THAGARD P. (2004) Causal inference in legal decision making. Explanatory coherence vs. Bayesian networks. *Applied Artificial Intelligence* **18,** 231–249.

THAGARD P. & VERBEURGT K. (1998) Coherence as constraint satisfaction. *Cognitive Science* **22,** 1–24.

TILLERS P. (2011) Trial by mathematics - reconsidered. *Law, Probability, and Risk* **10,** 167–173.

TILLERS P. & GOTTFRIED J. (2007) Case comment--United States v. Copeland, 369 F. Supp. 2d 275 (E.D.N.Y. 2005): A Collateral Attack on the Legal Maxim That Proof Beyond A Reasonable Doubt Is Unquantifiable? *Law, Probability and Risk* **5,** 135–157.

TRIBE L.H. (1971) Trial by Mathematics. Precision and Ritual in the Legal Process. *Harvard Law Review* **84,** 1329–1393.

TWINING W.L. (2006) Lawyers' stories. In *Rethinking evidence. Exploratory essays* (ed. W. L. Twining), pp. 286–331. Cambridge University Press, Cambridge.