Rethinking legal probabilism

Rafał Urbaniak

1 Scientific goal

As many miscarriages of justice indicate, scientific evidence is easily misinterpreted in court. This happens partially due to miscommunication between the parties involved, partially due to the usual probabilistic fallacies, but also because incorporating scientific evidence in the context of a whole case can be really hard. While probabilistic tools for piecemeal evaluation of scientific evidence and spotting probabilistic fallacies in legal contexts are quite well developed, the construction of a more general probabilistic model of incorporating such evidence in a wider context of a whole case, useful for theorizing about evidence evaluation and legal decision standards, remains a challenge. Legal probabilism (LP), for our purpose, is the view that this challenge can and should be met. This project intends to contribute to further development of this enterprise in a philosophically motivated manner.

The assessment of evidence in the court of law can be viewed from at least three perspectives: as an interplay of arguments, as an assessment of probabilities involved, or as an interaction of competing narrations. Each perspective presents an account of legal reasoning (Di Bello & Verheij, 2018; van Eemeren & Verheij, 2017). Individually, each of these strains has been investigated. The probabilistic approach, while being fairly mature, is still underdeveloped in light of various lines of criticism developed by the representatives of the other strains.

The **goal of this project** is to **develop of a probabilistic and yet narration-based modelling method of the interaction of various items of evidence and hypotheses and the resulting decisions in the court of law**. This will be achieved by accomodating important **insights provided by the critics of legal probabilism**. A crucial point of criticism is that the fact-finding process should be conceptualized as **a competition of narrations**. Another point comes from the argumentation theory framework: an adequate model should capture the structure of the arguments involved and the interplay between them. The key idea is that once **narrations are represented as bayesian networks, the argumentative structure becomes clear, various criteria on, features of and operations on narrations can be explicated in terms of corresponding properties of and operations on bayesian networks, and the interplay between arguments involved can be captured by the relations between bayesian networks or parts thereof**. The hypothesis is that such an improved framework will do better than the usual legal probabilisist's toolkit. Thus, the goals are three-fold:

1. Philosophical and conceptual improvement of legal probabilism by including a more holistic perspective and adversarial character of evidence assessment, which are typically absent from probabilistic approaches.
2. Formulation of a formal and computational probabilistic framework that incorporates features resulting from achieving goal 1. This will be done using Bayesian networks, hierarchical Bayesian models and imprecise probabilities. (**R** code capturing the technical features developed will be made openly available.)
3. Addressing the outstanding question of how evidence of different types can be aggregated and how adjudication should proceed in the presence of multiple competing narrations of what happened. The research in 1. and 2. will help to address this very practical pressing question.

Thus, the output will be a **unifying extended probabilistic model embracing key aspects of the narrative and argumentative approaches, susceptible to AI implementation.**

What the project will uniquely bring to the table is joining the familiarity with epistemological debates, familiarity with the details of evidence assessment in legal cases and technical skill to programmatically implement, simulate and test various theoretical moves.

2 Significance

2.1 State of the art

2.1.1 Legal probabilism

From among the three perspectives already mentioned, the probablistic approach will be my point of departure, for the following key reasons:

- The project is to be informed by and reflect on the actual practice of legal evidence evaluation, and much of scientific evidence in such contexts has probabilistic form.
- Probabilistic tools are fairly well-developed both for applications and within formal epistemology, reaching a state of fruition which should inspire deeper reflection.
- Statistical computing tools for such methods are available, which makes programming development and preliminary computational and data-driven evaluation of the ideas to be defended a viable enterprise.

Accoringly, the view in focus of this research is legal probabilism (LP)—an ongoing research program that comprises a variety of claims about evidence assessment and decision-making at trial. At its simplest, it comprises two core tenets: first, that the evidence presented at trial can be assessed, weighed and combined by means of probability theory; and second, that legal decision rules, such as proof beyond a reasonable doubt in criminal cases, can be explicated in probabilistic terms.

The early theorists of probability in the 17th and 18th century were as much interested in games of chance as they were interested in the uncertainty of trial decisions (Bernoulli, 1713; Daston, 1988; Franklin, 2001; Hacking, 1975). Bernoulli's prescient insights attained greater popularity in the 20th century amidst the law and economics movement (Becker, 1968; Calabresi, 1961; Posner, 1973). Finkelstein & Fairley (1970) gave one of the first systematic analyses of how probability theory, and Bayes' theorem in particular, can help to weigh evidence at trial. Lempert (1977) was one of the first to propose to use likelihood ratios for assessing the relevance of evidence. Such contributions fueled what has been called the New Evidence Scholarship, a rigorous way of studying the process of legal proof at trial (Lempert, 1986).

2.1.2 Challenges to New Evidence Scholarship

Tribe (1971) attacked what he called 'trial by mathematics', by listing well-known cases of misuse or probabilities in legal contexts and practical difficulties in assessing the probability of someone's criminal or civil liability, and pointing out the dehumanization of trial decisions that legal probabilism seems to propose. After Tribe, many argued that probabilistic models are either inadequate or unhelpful (Allen, 1986; Brilmayer, 1986; Cohen, 1986; Dant, 1988; Underwood, 1977). This negative trend has been somewhat mitigated by the discovery of DNA fingerprinting in the eighties and progress in forensic science in general, with the increasing role of quantitative evidence in the court of law (Kaye, 1986, 2010; Koehler, 1996; National Research Council, 1992; Robertson & Vignaux, 1995).

Skepticism about wider mathematical and quantitative models of legal evidence is still widespread among prominent legal scholars and practitioners (see, for example, Allen & Pardo, 2007). This is partially in light of conceptual difficulties extensively discussed in the literature, which arise when one wants to formulate a probabilistic decision criterion for the court of law. Imagine you are a trier of fact in a legal proceeding in which the defendant's guilt is identified as equivalent to a certain factual statement G and that somehow you succeeded in properly evaluating $\mathsf{P}(G|E)$—-the probability of G given the total evidence presented to you, E. One question that arises in such a situation is: when should you decide against the defendant? When is the evidence good enough? What we are after here is a condition Ψ, formulated in (primarily) probabilistic (and perhaps decision-theoretic) terms, such that the trier of fact, at least ideally, should accept any relevant claim A (including G) just in case $\Psi(A,E)$. One straightforward attempt might be to say: convict if $\mathsf{P}(G|E)$ is above a certain threshold, otherwise acquit (see, for example Laplace, 1814; Dekay, 1996; Kaye, 1979; Laudan, 2006).

This move, however, seems to be blocked by the so-called paradoxes of legal proof or puzzles of naked statistical evidence. Nesson (1979), Cohen (1981), and Thomson (1986) formulated scenarios in which, even if the probability of guilt or civil liability, based on the available evidence, is particularly high, a verdict against the defendant seems unwarranted. A variant of such a scenario—the gatecrasher paradox—goes as follows. Suppose our guilt threshold is high, say at 0.99. Consider the situation in which 1000 fans enter a football stadium, and 991 of them avoid paying for their tickets. A random spectator is tried for not paying. The probability that the

spectator under trial did not pay exceeds 0.99. Yet, intuitively, a spectator cannot be considered liable on the sole basis of the number of people who did and did not pay. While recently some doubt the relevance of abstract philosophical examples for the actual practice (Hedden, 2019; Ross, 2020), at least conceptual challenges remain.

Another conceptual problem is the so-called difficulty about conjunction. It arises, because intuitively there should be no difference between the trier's acceptance of A and B separately, and her acceptance of their conjunction, $A \wedge B$, that is, that $\Psi(A,E)$ and $\Psi(B,E)$ just in case $\Psi(A \wedge B,E)$. If $\Psi(H,E)$ is just the threshold criterion requiring that $\mathsf{P}(H|E)$ be sufficiently high, Ψ in general fails to satisfy this equivalence, as the probability of a conjunction generally can be lower than the probability of any of the conjuncts.

Arguably, these problems underscore a theoretical difficulty with probabilistic accounts of legal standards of proof. How to define them, or whether they should be even defined in the first place, remains contentious (Diamond, 1990; Horowitz & Kirkpatrick, 1996; Laudan, 2006; Newman, 1993; Walen, 2015). Judicial opinions offer different paraphrases, sometimes conflicting, of what these standards mean. In the last decade, philosophers have also joined the debate (for critical surveys see Redmayne, 2008; Gardiner, 2018)

At least *prima facie*, then, it seems that some conditions other than high posterior probability of liability have to be satisfied for the decision to penalize (or to find liable) to be justified. Accordingly, various informal notions have been claimed to be essential for a proper explication of judiciary decision standards (Haack, 2014; Wells, 1992). For instance, evidence is claimed to be insufficient for conviction if it is not *sensitive* to the issue at hand: if it remained the same even if the accused was innocent (Enoch & Fisher, 2015). Or, to look at another approach, evidence is claimed to be insufficient for conviction if it doesn't *normically support* it: if—given the same evidence—no explanation would be needed even if the accused was innocent (Smith, 2017). A legal probabilist needs either to show that these notions are unnecessary or inadequate for the purpose at hand, or to indicate how they can be explicated in probabilistic terms.

2.1.3 The alternative perspectives

More recently, alternative frameworks for modeling evidential reasoning and decision-making at trial have been proposed. They are based on inference to the best explanation (Allen, 2010; Hastie, 2019; Ho, 2019; Nance, 2019; Pardo & Allen, 2008; Schwartz & Sober, 2019), narratives and stories (Allen, 1986, 2010; Allen & Leiter, 2001; Clermont, 2015; Pardo, 2018; Pennington & Hastie, 1991a), and argumentation theory (Bex, 2011; Gordon, Prakken, & Walton, 2007; Walton, 2002). Those who favor a conciliatory stance have combined legal probabilism with other frameworks, offering preliminary sketches of hybrid theories (Urbaniak, 2018a; Verheij, 2014).

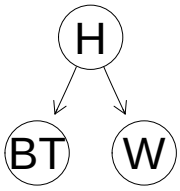
The main point of criticism is that legal proceedings are back-and-forth between opposing parties in which cross-examination is of crucial importance, reasoning goes not only evidence-to-hypothesis, but also hypotheses-to-evidence (Allen & Pardo, 2007; Wells, 1992) in a way that seems analogous to inference to the best explanation (Dant, 1988), which notoriously is claimed to not be susceptible to probabilistic analysis (Lipton, 2004), and the process is an interaction of multiple arguments that remain in fairly complicated relations. An informal philosophical account inspired by such considerations—The **No Plausible Alternative Story (NPAS)** theory (Allen, 2010)—is that the courtroom is a confrontation of competing narrations (Ho, 2008; Wagenaar, Van Koppen, & Crombag, 1993) offered by the sides, and the narrative to be selected should be the most plausible one. The view is conceptually plausible (Di Bello, 2013), and finds support in psychological evidence (Pennington & Hastie, 1991b, 1992).

It would be a great advantage of LP if it could model phenomena captured by the narrative approach. But how is the legal probabilist to make sense of them? From her perspective, the key disadvantage of NPAS is that it abandons the rich toolbox of probabilistic methods and takes the key notion of plausibility to be a primitive notion which should be understood only intuitively. It would be even better for LP if the incorporation of such insights could lead to the resolution of the already mentioned conceptual difficulties.

2.1.4 Bayesian networks as a tool for legal probabilism

The idea that Bayesian networks can be used for probabilistic reasoning in legal fact-finding started gaining traction in late eighties and early nineties (Edwards, 1991), and it found its way to nowadays standard textbooks on the topic (Fenton & Neil, 2018a; Taroni, Biedermann, Bozza, Garbolino, & Aitken, 2014).
A Bayesian network comprises two components: first, a directed acyclic graph of relations of dependence

(represented by arrows) between variables (represented by nodes); second, conditional probability tables. Consider the graphical component first. The graph is acyclic because the arrows connecting the nodes do not form loops. As an illustration, let H be the claim that the suspect committed the murder, BT the presence of a blood type B match with a crime scene stain, and W the fact that an eyewitness observed the suspect near the scene around the time of the crime. The graphical component of the Bayesian network would look like this.



The *ancestors* of a node X are all those nodes from which we can reach X by following the arrows going forwards. The *parents* of a node X are those for which we can do this in one step. The *descendants* of X are all which can be reached from X by following the arrows going forward. The *children* are those for which we can do this in one step. In the example, H is the parent (and ancestor) of both W and BT, which are its children (and descendants). There are no non-parent ancestors or non-children descendants.

The variables, which are represented by nodes and are connected by arrows, stand in relation of probabilistic dependence. To describe these relations, the graphical model is accompanied by conditional probability tables. For instance, they can look as follows (the blood type frequency estimate is realistic (Lucy, 2013), and so are the conditional probabilities for the eyewitness identification, although for complications about assessing eyewitness testimony see Wixted & Wells (2017) and Urbaniak, Kowalewska, Janda, & Dziurosz-Serafinowicz (2020)).

	H=murder	H=no.murder
W=seen	.7	.4
W=not.seen	.3	.6

	H=murder	H=no.murder
BT=match	1	.063
BT=no.match	0	.937

and a prior for the states of the root nodes (here, say, $\mathsf{P}(H = murder) = .01$ and $\mathsf{P}(H = no.murder) = 0.99$).

While the Bayesian network above—comprising a directed acyclic graph along with probability tables—is simple, a correct intuitive assessment of the probability of the hypothesis given the evidence is already challenging. The reader is invited to try to estimate intuitively the probability that the defendant committed the murder (H=murder) given the following states of the evidence:

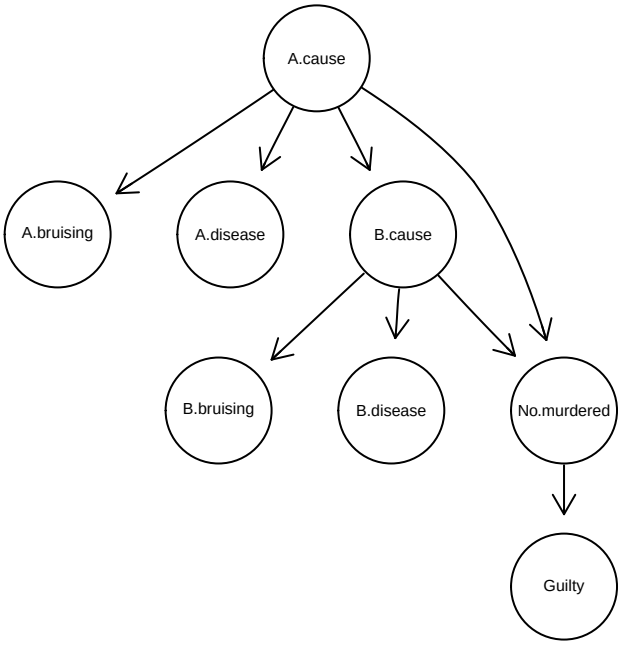
- The suspect's blood type matches the crime stain but information about the witness is unavailable.
- The suspect's blood type matches the crime stain but the witness says they did not see the suspect near the crime scene.
- The suspect's blood type matches the crime stain and the witness says they saw the suspect near the crime scene.

	H=murder
BT=match,W=?	.138
BT=match,W=not.seen	.074
BT=match, W=seen	.219

Already at this level of complexity, calculations by hand become cumbersome. In contrast, software for Bayesian networks will easily give the results visible on the left. Perhaps surprisingly, the posterior probability of H is about .22 even when both pieces of evidence are incriminating (BT=match, W=seen).

In a similar vein, fairly simple graphical patterns (called *idioms*) often recur while modeling the relationships between evidence and hypotheses. Complex graphical models can be created by combining these basic patters in a modular way. Discussion of general methods for Bayesian network constructions can be found in (Bovens & Hartmann, 2004; Friedman, 1974; Hepler, Dawid, & Leucari, 2007; Neil, Fenton, & Nielson, 2000) and general idioms are discussed in (Fenton, Neil, & Lagnado, 2013).

Some attempts have been made to use Bayesian networks to weigh and assess complex bodies of evidence consisting of multiple components. On one hand, we have serious reconstructions of real complex cases. Kadane & Schum (2011) made one the first attempts to model an entire criminal case, Sacco & Vanzetti from 1920, using probabilistic graphs. Here is another, more recent, example by Fenton & Neil (2018b), who constructed a Bayesian network for the famous Sally Clark case.



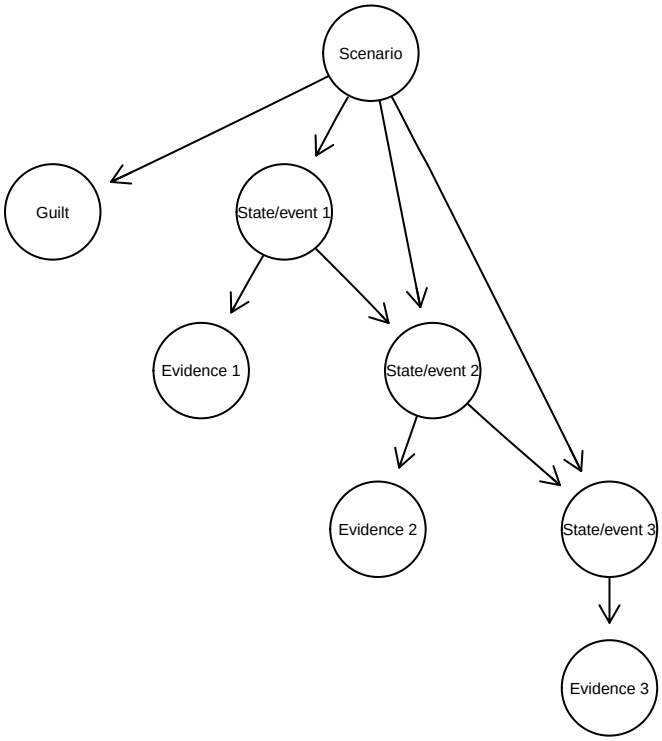
The arrows depict relationships of influence between variables. Whether Sally Clark's sons, call them *A* and *B*, died by SIDS or murder (A.cause and B.cause) influences whether signs of disease (A.disease and B.disease) and bruising (A.bruising and B.bruising) were present. Since son A died first, whether A was murdered or died by SIDS (A.cause) influences how son B died (B.cause). How the sons died determines how many sons were murdered (No.murdered), and how many sons were murdered decides whether Sally Clark is guilty (guilty).

Evidence (cumulative)	P(Clark guilty)
A bruising	.2887
A no signs of disease	.3093
B bruising	.6913
B no signs of disease	.7019

In the original calculation, the prior probability of Guilty = Yes should be .0789. After taking into account the incriminating evidence presented at trial, such as that there were signs of bruising but no signs of a preexisting disease affecting the children, the posterior probabilities are as in the table on the left.

The literature contains examples of more general methodological reflection on the use of BNs for modeling whole cases. The main idea is that once all the pieces of evidence and claims are represented as nodes, one should use the *scenario idiom* to model complex hypotheses, consisting of a sequence of events organized in space and time: a scenario (Vlek, Prakken, Renooij, & Verheij, 2014). A discussion of modelling crime scenarios by means of graphical devices mixed with probabilities can be also found in the work of Shen, Keppens, Aitken, Schafer, & Lee (2007)}, Bex (2011), Bex (2015) and Verheij (2017). See also the survey by Di Bello & Verheij (2018). Dawid & Mortera (2018) give a treatment of scenarios in terms of BNs.

A graphical model that uses the scenario idiom would consist of the following components: first, nodes for the states and events in the scenario, with each node linked to the supporting evidence; second, a separate scenario node that has states and events as its children; finally, a node corresponding to the ultimate hypothesis as a child of the scenario node. Such a
model could look like this:



Note that the scenario node unifies the different events and states. Because of this unifying role, increasing the probability of one part of the scenario (say State/event 2) will also increase the probability of the other parts (State/event 1 and State/event 3). This is intended to capture the idea that the different components of a scenario form an interconnected sequence of events.

One challenge that this strategy is supposed to help with is the question of how to make sense of the notion of the coherence of a scenario as different from its probability given the evidence. On this approach (Vlek, 2016; Vlek, Prakken, Renooij, & Verheij, 2013; Vlek et al., 2014; Vlek, Prakken, Renooij, & Bart Verheij, 2015), coherence is identified with the prior probability of the scenario node.

Another challenge that the framework is supossed to meet is the question of how to formally represent reasoning with multiple scenarios on the table. On this approach (called scenario merging), given a class of narrations, all the nodes used in some of the separate BNs are to be used to build one large BN, and separate scenario nodes are to be added to it, so that one BN supposedly represents multiple scenarios at once.

A somewhat alternative approach to representation of and reasoning with multiple scenarios has been developed by Neil, Fenton, Lagnado, & Gill (2019). They correctly criticize (Urbaniak, 2018a) where I only sketch some theoretical moves in a second-order language towards the probabilistic modelling of the narrative approach. The critics point out the paper makes no connection to BNs and so it "fails to offer a convincing and operational means to structure and compare competing narratives." This is a fair assessment of the limits of what I have achieved so far. They propose to represent separate narrations in terms of separate BNs, and to deploy bayesian model comparison and averaging as a tool for reasoning with multiple scenarios. That is, Bayes Theorem with hypotheses as models (BNs), yields:

$$\mathsf{P}(M = m_i|E) = \frac{\mathsf{P}(E|M = m_i)\mathsf{P}(M = m_i)}{\sum_{i=1}^{n} \mathsf{P}(E|M = m_i)\mathsf{P}(M = m_i)} \tag{1}$$

Then, assuming equal priors, models with higher likelihoods will have higher posterior probabilities, and the most plausible model will be the one with the highest posterior (that is, with equal priors, with highest likelihood). Alternatively, they propose averaging the predictions for a given variable φ by taking the ensemble model:

$$\mathsf{P}(\varphi|E) = \sum_{i=1}^{n} \mathsf{P}(\varphi|M = m_i, E)\mathsf{P}(M = m_i|E) \tag{2}$$

where the priors are either equal or are identified with the posterior of the models given the evidence, and those posteriors are to be calculated assuming equal priors.

2.2 Pioneering nature of the project

2.2.1 Points of disagreement

Here are the key reasons why I am convinced the scenario node approach is not satisfactory:

A. The use of a scenario idiom is problematic. Adding a parent node by *fiat* without any good reasons to think the nodes are connected other than being part of a single story, introduces probabilistic dependencies between the elements of a narration. Merely saying that, say, the defenendant made jointly some claims is not a good reason to assume they are probabilistically dependent.

B. Another problem results from the identification of prior probability with coherence. This does not add up intuitively. After all, it is quite coherent with my views that if I win a lottery, I'll buy a large house in Auckland and move there, while both the prior and the posterior given the total available evidence of this scenario are rather low.

C. In general, the legal probabilistic approach to coherence is very simple and fails to engage with rich philosophical literature exactly on this topic (Douven & Meijs, 2007; Fitelson, 2003a, 2003b; Glass, 2002; Meijs & Douven, 2007; Olsson, 2001; Shogenji, 1999), including a long list of counterexamples to the existing proposals and desiderata that a probabilistic coherence measure should satisfy (Akiba, 2000; Bovens & Hartmann, 2004; Crupi, Tentori, & Gonzalez, 2007; Koscholke, 2016; Merricks, 1995; Schippers & Koscholke, 2019; Shogenji, 1999, 2001, 2006; Siebel, 2004, 2006).

D. The merging procedure with scenario nodes assumes that for the nodes that are common to the networks to be merged, both the directions of the arrows in the DAGs and the conditional probability tables are the same across different narrations. This is suboptimal. Different sides in court might construe causal dependencies differently, and even if they agree about the direction of an arrow, they might disagree about the probability table associated with it. Even a single side might consider different scenarios with different probabilities, say, when there is some uncertaintly inolved in the probability assignment itself.

Here are the key limitations of the approach proposed by Neil et al. (2019):

E. The assumption of equal priors is highly debatable. For one thing, this approach would render prior probabilities quite sensitive to the choice of hypotheses and thus potentially arbitrary. In addition, this approach seems particularly unsuitable for criminal cases. If the only two hypotheses/models on the table ultimately say "the defendant is guilty" and "the defendant is innocent", the prior probability of each would be 50%. But defendants in criminal cases, however, should be presumed innocent until proven guilty. A 50% prior probability of guilt seems excessive. Some (Williamson, 2010) try to defend a variant of the principle of indifference by reference to informational entropy, and a proposal along this line has been used in practical recommendation by expert committees (ENFSI Expert Working Group Marks Conclusion Scale Committee, 2006). However, this attempt has been sensibly criticized by Biedermann, Taroni, & Garbolino (2007). The question remains, what should the proper application of informational entropy in the context of BN selection and averaging look like, given that information entropy considerations independently are in epistemologically decent standing? Importantly, whatever conclusions in such contexts are epistemologically justified, how do they square with the presumption of innocence? Some take the presumption to mean that the prior probability of guilt should be set to a small value (Friedman, 2000; Friedman, Allen, Balding, Donnelly, & Kaye, 1995), but it is not clear whether this interpretation can be justified on epistemological or decision-theoretic grounds.

F. More recent models rely on relevant background information, for example, geographical information about people's opportunities to commit crimes [Fenton, Lagnado, Dahlman, & Neil (2019). But even if these models are successful in giving well-informed assessments of prior probabilities any evidence-based assessment of prior probabilities, they are likely to violate existing normative requirements of the trial system (Dahlman, 2017; Engel, 2012; Schweizer, 2013). For instance, if the assessment of prior probabilities relies on demographic information, people who belong to certain demographic groups will be regarded as having a higher prior probability of committing a wrong than others. This is what a well-informed assessment should amount to. Yet, if some people's priors are higher than other people's priors, it will be easier to convict or find liable those who are assigned higher priors, even if the evidence against them is the same as the evidence against those assigned lower priors. This outcome can be seen as unfair (Di Bello & O'Neil, 2020). The question remains: what procedure of choosing the priors both is justified by epistemological considerations and does not generate tension with fairness considerations?

G. Model selection based on likelihood (given equal priors) or posterior model probabilities in general (if priors are not assumed to be equal) boils down to a variant of the threshold view, and so all the difficulties with the threshold view apply.

H. Model averaging in the proposed form boils down to taking a weighted average of the probabilities provided by the models (weighted linear pooling). However, there is a rich literature on the difficulties that linear pooling runs into (see the surveys in Dietrich & List, 2016; Franz Dietrich & List, 2017a, 2017b). One problem is that the method satisfies the unanimity assumption: whenever all models share a degree of belief in a claim, this is exactly the output degree for that belief. But clearly, a claim can receive additional boost from multiple agents with different pieces of evidence agreeing on something (for instance, in witness corroboration). Another problem is that linear pooling does not preserve probabilistic independence (List & Pettit, 2011): even if all models agree that certain nodes are independent, they might end up being dependent in the output. There is also a variety of impossibility theorems in the neighborhood (Gallow, 2018).[1]

2.2.2 Strategy and novelty

First, we need to represent narrations with their argumentative and dynamic structure in probabilistic terms. Once this is done, the resulting rich structure can be used to work out a better explication of the notion of coherence. Once this is done, various methods of dealing with multiple BNs, be it ensemble methods or some model selection methods, should be put into place, by explaining in a principled way which of these methods should be used in what context and why. One key novelty is in the formulation and BN implementation of selection criteria that so far have only been discussed informally in philosophical literature. Once this explication is obtained, further investigation of the role these criteria should play and their principled justification will proceed. I cover these steps in more detail now.

Representation. I will use BNs taken separately without scenario nodes to represent various narrations. Crucially, I will not assume the conditional probability tables or directions of edges are the same across the BNs, thus allowing for more realistic flexibility. To be able to accommodate insights provided by NPAS and other critics of LP, I will add another layer of information: for each BN one needs to specify a set of binary nodes such that a certain combination of their states counts as a narration, and a set of evidence nodes, which are supposed to support this narration.

Dynamic BNs If what we are modeling is cross-examination, averaging does not seem to be the right way to go. To model cross-examination, we need to take the argumentative approach seriously and to be able to model relations such as "undercutting" and "rebutting". And these relations can be modelled by adding or removing nodes or arrows in the network in a suitable manner. But this means we might have to consider another dimension: BNs changing through time in light of other BNs. How to make sense of this formally so that the result makes sense philosophically remains an open question.

BN-based coherence. I have developed a coherence measure that diverges from the known candidates in three important respects: (1) It is not a function of a probabilistic measure and a set of propositions alone, because it is also sensitive to the selection and direction of arrows in a Bayesian Network representing an agent's credal state. (2) Unlike in the case of quite a few coherence measures, it is not obtained by taking a mean of some list of intermediate values (such as confirmation levels between subsets of a narration). It is sensitive also to the variance and the minimal values of the intermediate values. (3) The intermediate values used are not confirmation levels, but rather expected and weighted confirmation levels. Preliminary tests on existing philosophical counterexamples suggests the performance of the measure is much better than the existing coherence measures. Now, it needs to be deployed (implemented in **R** for BNs) and properly tested on real-life cases discussed in the LP literature.

Divide and conquer. In fact, dealing with multiple models is difficult in this context. On one hand, many

[1] Here is a nice example. It turns out you can't at the same time hold the following:

$$\mathsf{P}(A = B) < 1 \tag{3}$$

$$\mathsf{P}(r|A = a) = a \tag{4}$$

$$\mathsf{P}(r|B = b) = b \tag{5}$$

$$\forall a,b\,\mathsf{P}(r|A = a, B = b) = \alpha a + \beta b \tag{6}$$

This means that that it is impossible that two models A and B can disagree, we trust each of them separately if we only learn about one model, and we take a weighted average if we learn about both.

machine learning methods are not available. For instance, one cannot evaluate models in terms of their performance with respect to the data. Whenever you want to use resampling methods (such as cross-validation), or some information criterion scoring (suchs as Akaike Information Criterion), you need to have a dataset with multiple datapoints to start with, and such datasets are usually not available (and often conceptually unimaginable) for the problems typically faced in the court of law. On the other hand, averaging often doesn't make sense either. After all, often no epistemological or decision-related progress might be gained based on averaging the prosecutor's and the defendant's stories. I propose that ensemble methods should be deployed for multiple narration variants available from one side (as in when, say, the prosecution story comes with uncertainty about the direction of an arrow or about a particular probability table), but selection methods should be used when final decision is to be made between narrations proposed by the opposing sides.

Ensemble methods. One question that arises is whether the general concerns about linear pooling arise for such limited applications. If not, the remaining concern is what priors should be used. In light of the controversial nature of equal priors, I plan to study the consequences of rescaling coherence scores (already mentioned) to constitute model priors. The idea is that given that narrations are to be developed by the sides themselves, taking coherence of their narration as determining the prior might be more fair than using equal priors or relying on geographical or population statistics. If yes, perhaps some other methods boiling down to a variant of sensitivity analysis can be deployed: look at all BNs corresponding to some variant of the narration of one of the sides, find the strongest and the weakest one, and these give you a range of possible outcomes.

Selection criteria. The so-called New Legal Probabilism (NLP) is an attempt to improve on the underspecificity of NPAS (Di Bello, 2013). While still being at most semi-formal, the approach is more specific about the conditions that a successful accusing narration is to satisfy for the conviction beyond reasonable doubt to be justified. Di Bello identifies four key requirements that a successful convicting narration should satisfy:

(Evidential support)	The defendant's guilt probability on the evidence should be sufficiently supported by the evidence, and a successful accusing narration should explain the relevant evidence.
(Evidential completeness)	The evidence available at trial should be complete as far as a reasonable fact-finders' expectations are concerned.
(Resiliency)	The prosecutor's narrative, based on the available evidence, should not be susceptible to revision given reasonably possible future arguments and evidence.
(Narrativity)	The narrative offered by the prosecutor should answer all the natural or reasonable questions one may have about what happened, given the content of the prosecutor's narration and the available evidence.

Prima facie, it is far from obvious that such conditions are susceptible to a Bayesian networks explication. However, I have already developed a more expressive probabilistic framework (call it Narration-Friendly Probabilism (NFP)) capable of expressing such features within a formalized higher-order language (Urbaniak, 2018b). On NFP, the notion of narration is quite wide: narrations not only contain factual statements about what happened, but also claims about evidence, about narrations, about relations between evidence and various parts of various narrations etc. I extend the basic propositional language with propositional operators N_i and E corresponding to "*... is part of narration i*" and "*... is part of the evidence*," and model narrations as finite sets of sentences from this language. Due to this intuitive move, many important aspects of narrations normally discussed only informally, similar to those discussed by Di Bello, become expressible in terms of probabilistic measures for such a formal language. Let's very briefly gesture towards a few examples:

- A defending narration explains a piece of evidence e just in case if there is an attacking narration whose posterior is raised conditional on e, the probability of e conditional on this defending narration is above the negligibility threshold.
- An attacking narration misses some evidence just in case there are some statements not in the evidence set such that the probability of the claim that at least one of them is part of evidence conditional on the existing evidence ($\{\varphi|\varphi \in$ Evidence$\}$), its description ($\{E\varphi|\varphi \in$ Evidence$\}$), and on this attacking narration is above the strong plausibility threshold.
- A narration contains gaps just in case there are some claims which are not part of it, but conditional on the content and the description of this narration and the evidence available, their disjunction is strongly plausible, and it is strongly plausible conditional on the content (but not on the description) of this narration

and on evidence that at least one of these claims is part of the narration.
- A narration is dominant just in case it doesn't miss any evidence, it doesn't contain any gap, and in light of all available information and evidence it is at least as likely any other accusing narration, and is strongly plausible.

While threshold- or likelihood-ratio-based selection criteria for models are unlikely to succeed, as already discussed, I am convinced the criteria formulated in philosophical terms in (Di Bello, 2013) and in higher-order terms in (Urbaniak, 2018a) are in better standing. The key hypothesis is that they can be recast in terms of properties of BNs and that the existing BN programming tools can be extended to implement testing for these criteria. This will make them susceptible to programmatic implementation and further study by means of computational methods. The hope is that on one hand, they will do better than the existing proposals, and where they fail, further insights can be gained by studying the reasons behind such failures.

3 Work plan

Stage 1 Philosophical & formal unification	Obtain a unifying extended probabilistic framework by incorporating further insights from philosophical and psychological accounts of legal narrations, and from the argumentation approach. Defend its philosophical plausibility. (6 months)
Stage 2 AI implementation	Develop Bayesian Network Methods for the obtained formal framework, so that the insights from the argumentation approach and informal epistemology, mediated through it, can be incorporated in AI tools. (12 months)
Stage 3 Case studies	Evaluate the developed framework and AI tools by conducting case studies from its perspective. (6 months)
Stage 4 Back to challenges & output	Investigate the extent to which the new framework helps to handle the issues raised in points A.-H., finalize the book. (12 months)

The planned publication output is as follows:

- Stage 1 will result in one philosophical paper published in an academic journal such as *Synthese*, *Mind* or *Ratio Juris*. Working title: *Why care about narration selection principles?*
- Stage 2 will lead to one technical paper published in a journal such as *IfCoLog Journal of Logics and their Applications*, *Law, Probability and Risk* or *Artificial Intelligence and Law*. Working title: *Implementation of narration assessment criteria in Bayesian Networks with **R**.*
- Stage 3 will lead to a publication of one paper on how the formal framework handles case studies and a further paper on how the developed AI tools handle real-life situations in journals such as *Artificial Intelligence* or *Argument & Computation*. Working titles: *Rethinking the famous BN-modeled cases within the narration framework* and *BNarr, an **R** package to model narrations with Bayesian Networks*.
- Throughout the whole project I plan to cooperate with Marcello Di Bello. Over the last year we co-authored the Stanford Encyclopedia of Philosophy entry on Legal Probabilism. We decided to continue our fruitful cooperation. For over two months we have been working on a book proposal to be submitted to Oxford University Press exactly on the issues to be studied in this research project. Marcello Di Bello is an excellent philsopher with extensive research experience in the philosophy of legal evidence, and he would bring his expertise to the table when working both on the book and on the papers, whereas I would be focused on the technical aspects and the underlying formal philosophy. During the last year of the research **I plan a six months' stay at Arizona University**, to work in person with Di Bello on finalizing the book that presents the results with special focus on issues investigated in Stage 4.

The tentative list of planned chapters is as follows (two of them already exist as sample chapters for the book proposal submission):

I Legal probabilism and its foes
- 1 The emergence of legal probabilism
  - 1.1 Famous cases
  - 1.2 Probabilistic evidence
  - 1.3 Trial by mathematics
  - 1.4 Some history
- 2 A skeptical perspective
  - 2.1 The difficulty about conjunction
  - 2.2 The complexity objection
  - 2.3 The problem of corroboration
  - 2.4 The problem of artificial precision
  - 2.5 Naked statistical evidence
  - 2.6 The problem of priors
  - 2.7 The reference class problem
  - 2.8 Non-probabilistic perspectives

II Evidence assessment
- 3 Bayes' Theorem and the usual fallacies
  - 3.1 Assuming independence
  - 3.2 The prosecutor's fallacy
  - 3.3 Base rate fallacy
  - 3.4 Defense attorney's fallacy
  - 3.5 Uniqueness fallacy
  - 3.6 Case studies
- 4 Complications and caveats
  - 4.1 Complex hypotheses and complex bodies of evidence
  - 4.2 Source, activity and offense level hypotheses
  - 4.3 Where do the numbers come from?
  - 4.4 Modeling corroboration
  - 4.5 Stories, explanations and coherence
- 5 Likelihood Ratios and Relevance
  - 5.1 Likelihood ratio as a measure of evidence strength
  - 5.2 The risk of false positive and its impact
  - 5.3 Hypothesis choice
  - 5.4 Levels of hypotheses and the two-stain problem
  - 5.5 Relevance and the small-town murder scenario
  - 5.6 The cold-hit confusion
  - 5.7 Likelihood ratio and cold-hit DNA matches
- 6 Bayesian Networks
  - 6.1 Bayesian networks to the rescue
  - 6.2 Legal evidence idioms
  - 6.3 Scenario idioms
  - 6.4 Modeling relevance
  - 6.5 Case study: Sally Clark
  - 6.6 DNA evidence
- 7 Corroboration
  - 7.1 Boole's formula and Cohen's challenge
  - 7.2 Modeling substantial rise in case of agreement
  - 7.3 Ekelöf's corroboration measure and evidentiary mechanisms
  - 7.4 General approach with multiple false stories and multiple witnesses
- 8 Coherence
  - 8.1 Existing probabilistic coherence measures
  - 8.2 An array of counterexamples
  - 8.3 Coherence of structured narrations with Bayesian networks
  - 8.4 Application to legal cases
- 9 New legal probabilism
  - 9.1 Desiderata
  - 9.2 A probabilistic framework for narrations
  - 9.3 Probabilistic explications of the desiderata
  - 9.4 Bayesian network implementation

III Trial Decisions
- 10 The functions of the proof standards
  - 10.1 Conceptual desiderata
  - 10.2 Protecting defendants
  - 10.3 Error reduction and error distribution/allocation
  - 10.4 Dispute resolution and public deference
  - 10.5 Justification and answerability
- 11 Standards of proof
  - 11.1 Legal background
  - 11.2 Probabilistic thresholds
  - 11.3 Theoretical challenges
  - 11.4 Specific narratives
  - 11.5 The comparative strategy
  - 11.6 The likelihood strategy
  - 11.7 Challenges (again)
  - 11.8 Probabilistic thresholds revised
  - 11.9 Bayesian networks and probabilistic standard of proof
- 12 Accuracy and the risk of error
  - 12.1 Minimizing expected costs
  - 12.2 Minimizing expected errors
  - 12.3 Expected v. actual errors
  - 12.4 Competing accounts of the risk of error
  - 12.5 Bayesian networks and the risk of error
- 13 Fairness in trial decisions
  - 13.1 Procedural v. substantive fairness
  - 13.2 Competing measures of substantive fairness
  - 13.3 Bayesian networks and fairnesss
- 14 Alternative accounts and legal probabilism
  - 14.1 Baconian probability
  - 14.2 Relative Plausibility
  - 14.3 Arguments
  - 14.4 Sensitivity
  - 14.5 Normic Support
  - 14.6 Justification/foundherentism
  - 14.7 Completeness
  - 14.8 Relevant alternatives
  - 14.9 Knowledge
- 15 Conclusions

Apart from publications, the results will be presented at various conferences devoted to legal reasoning. These include the yearly conferences of the *International Association for Artificial Intelligence and Law* and of the *Foundation for Legal Knowledge Based Systems (JURIX)*, and more general conferences gathering formal philosophers, so that the research is inspired by interaction not only with legal evidence scholars, computer scientists, but also philosophers. I am also already an invited speaker at the upcoming "Probability and Proof"

conference that will be part of an international conference on the philosophy of legal evidence ("The Michele Taruffo Girona Evidence Week") in Girona (Spain) May 23-27 2022.

4 Methodology

Standard arguments for the legitimacy of Bayesianism[2] deploy usually rather abstract pieces of reasoning to the effect that if one's degrees of beliefs satisfy certain conditions, they also have to satisfy the probabilistic requirements. My approach to thinking about the plausibility of Bayesian epistemology is rather unlike such approaches. Instead, I prefer the *proof-of-the-pudding* methodology. I am convinced that an important part of the philosophical assessment of the Bayesian research program has to do with its achievements or failures in contributing to debates in philosophy which are not themselves debates about the status of Bayesianism itself. In particular, it would be great news if insights from Bayesian epistemology could be used to further development of forensic AI and deepening our understanding of judiciary decision making.

What the project will uniquely bring to the table is joining the familiarity with epistemological debates on the nature of coherence (which legal probabilists like Vlek or Fenton ignore or are unaware of), familiarity with the details of evidence assessment in legal cases (which formal epistemologists such as Fitelson ignore) and technical skill to programmatically implement, simulate and test various theoretical moves.

I will be using four methods: (a) informal conceptual analysis; (b) formal conceptual analysis; (c) computational metods (R simulations, etc.); and (d) case studies. I will rethink and model the existing impliementations of whole-case-scale-BNs in legal evidence from the perspective of the new framework and reconstruct cases which extensively use probabilistic reasoning, but for which BNs have not been yet proposed. Both the literature already listed and many textbooks on quantitative evidence in forensics are great sources of such cases.

A larger initiative involving reconstructing various cases using different representation methods and comparing the representations, called *Probability and statistics in forensic science*, took place at the Isaac Newton Institute for Mathematical Sciences. My approach will be in the same vein.

I have extensive experience in analytic philosophy, conceptual analysis, philosophical logic (including non-monotonic logics, such as adaptive logics which have been developed at Ghent University, where I spent quite a few years), and probabilistic and decision-theoretic methods as deployed in philosophical contexts. I also have teaching and publishing record that involves statistical programming in the **R** language (my sample programming projects can be visited at https://rfl-urbaniak.github.io/menu/projects.html), and so am also competent to develop AI implementations and tests of the ideas to be developed.

One research risk is that it will turn out that some of the informal requirements cannot be spelled out in probabilistic terms, or expressed in terms of properties of Bayesian networks. In such an event, I will study the reasons for this negative result. It might happen that there are independent reasons to abandoned a given condition, or it might be the case that probabilistic inexplicability of a given condition is an argument against the probabilistic approach. Either way, finding out which option holds and why would also lead to a deeper understanding of the framework and lead to a publication in academic journals. Another research risk is that the case studies will show that other methods are more efficient or transparent. This would itself constitute a result that could be used to further modify the framework so that its best aspects could be preserved, while the disadvantages discovered during case studies avoided.

5 References

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

Allen, R. J. (1986). A reconceptualization of civil trials. *Boston University Law Review*, *66*, 401–437.

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

Allen, R. J., & Leiter, B. (2001). Naturalized epistemology and the law of evidence. *Virginia Law Review*, *87*(8), 1491–1550. JSTOR.

Allen, R., & Pardo, M. (2007). The problematic value of mathematical models of evidence. *The Journal of Legal Studies*, *36*(1), 107–140. JSTOR.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, *76*, 169–217. Springer.

[2] See for example (Earman, 1992; Urbach & Howson, 1993) for an early yet fairly comprehensive survey, or (Pettigrew, 2011) for a discussion of more recent contributions. See also (Bovens & Hartmann, 2004; Bradley, 2015; Swinburne, 2001).

Bernoulli, J. (1713). *Ars conjectandi*.

Bex, F. (2015). An integrated theory of causal stories and evidential arguments. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law - ICAIL '15* (pp. 13–22). San Diego, California: ACM Press.

Bex, F. J. (2011). *Arguments, stories and criminal evidence: A formal hybrid theory*. Law and philosophy library. Dordrecht ; New York: Springer.

Biedermann, A., Taroni, F., & Garbolino, P. (2007). Equal prior probabilities: Can one do any better? *Forensic Science International*, *172*(2-3), 85–93. Elsevier BV. Retrieved from https://doi.org/10.1016/j.forsciint.2006.12.008

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

Bradley, D. (2015). *A critical introduction to formal epistemology*. Bloomsbury Publishing.

Brilmayer, L. (1986). Second-order evidence and bayesian logic. *Boston University Law Review*, *66*, 673–691.

Calabresi, G. (1961). Some thoughts on risk distribution and the law of torts. *Yale Law Journal*, *70*, 499–553.

Clermont, K. M. (2015). Trial by Traditional Probability, Relative Plausibility, or Belief Function? *Case Western Reserve Law Review*, *66*(2), 353–391.

Cohen, J. L. (1981). Subjective probability and the paradox of the Gatecrasher. *Arizona State Law Journal*, 627–634.

Cohen, J. L. (1986). Twelve questions about Keynes's concept of weight. *British Journal for the Philosophy of Science*, *37*(3), 263–278.

Crupi, V., Tentori, K., & Gonzalez, M. (2007). On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues. *Philosophy of Science*, *74*(2), 229–252.

Dahlman, C. (2017). Determining the base rate for guilt. *Law, Probability and Risk*, *17*(1), 15–28.

Dant, M. (1988). Gambling on the truth: The use of purely statistical evidence as a basis for civil liability. *Columbia Journal of Law and Social Problems*, *22*, 31–70. HeinOnline.

Daston, L. (1988). *Classical probability in the enlightenment*. Princeton University Press.

Dawid, A. P., & Mortera, J. (2018). Graphical models for forensic analysis. In *Handbook of graphical models* (pp. 491–514). CRC Press.

Dekay, M. L. (1996). The difference between Blackstone-like error ratios and probabilistic standards of proof. *Law and Social Inquiry*, *21*, 95–132.

Diamond, H. A. (1990). Reasonable doubt: To define, or not to define. *Columbia Law Review*, *90*(6), 1716–1736.

Di Bello, M. (2013). *Statistics and probability in criminal trials* (PhD thesis). University of Stanford.

Di Bello, M., & O'Neil, C. (2020). Profile evidence, fairness and the risk of mistaken convictions. *Ethics*, *130*(2), 147–178.

Di Bello, M., & Verheij, B. (2018). Evidential reasoning. In *Handbook of legal reasoning and argumentation* (pp. 447–493). Springer.

Dietrich, F., & List, C. (2016). Probabilistic opinion pooling. In A. Hajek & C. Hitchcock (Eds.), *Oxford handbook of philosophy and probability*. Oxford University Press.

Dietrich, F., & List, C. (2017a). Probabilistic opinion pooling generalized. Part one: General agendas. *Social Choice and Welfare*, *48*(4), 747–786. Springer.

Dietrich, F., & List, C. (2017b). Probabilistic opinion pooling generalized. Part two: The premise-based approach. *Social Choice and Welfare*, *48*(4), 787–814. Springer.

Douven, I., & Meijs, W. (2007). Measuring coherence. *Synthese*, *156*(3), 405–425. Springer Science; Business Media LLC. Retrieved from https://doi.org/10.1007/s11229-006-9131-z

Earman, J. (1992). *Bayes or bust? A critical examination of bayesian confirmation theory*. Cambridge: MIT Press.

Edwards, W. (1991). Influence diagrams, bayesian imperialism, and the collins case: An appeal to reason. *Cardozo Law Review*, *13*, 1025–1074.

ENFSI Expert Working Group Marks Conclusion Scale Committee. (2006). Conclusion scale for shoeprint and toolmarks examinations. *Journal of Forensic Identification*, *56*, 255–280.

Engel, C. (2012). Neglect the Base Rate: It's the Law! *Preprints of the Max Planck Institute for Research on Collective Goods*, *23*.

Enoch, D., & Fisher, T. (2015). Sense and sensitivity: Epistemic and instrumental approaches to statistical evidence. *Stan. L. Rev.*, *67*, 557–611. HeinOnline.

Fenton, N., Lagnado, D., Dahlman, C., & Neil, M. (2019). The opportunity prior: A proof-based prior for criminal cases. *Law, Probability and Risk*, [online first].

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

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

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A General Structure for Legal Arguments About Evidence Using Bayesian Networks. *Cognitive Science*, *37*(1), 61–102.

Finkelstein, M. O., & Fairley, W. B. (1970). A Bayesian approach to identification evidence. *Harvard Law Review*, *83*(3), 489–517.

Fitelson, B. (2003a). A Probabilistic Theory of Coherence. *Analysis*, *63*(3), 194–199.

Fitelson, B. (2003b). Comments on jim franklin's the representation of context: Ideas from artificial intelligence (or, more remarks on the contextuality of probability). *Law, Probability and Risk*, *2*(3), 201–204. Oxford Univ Press.

Franklin, J. (2001). *The science of conjecture: Evidence and probability before pascal*. John Hopkins University Press.

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy*, *71*, 5–19.

Friedman, R. D. (2000). A presumption of innocence, not of even odds. *Stanford Law Review*, *52*(4), 873–887.

Friedman, R. D., Allen, R. J., Balding, D. J., Donnelly, P., & Kaye, D. H. (1995). Probability and proof in State v. Skipper: An internet exchange. *Jurimetrics*, *35*(3), 277–310.

Gallow, J. (2018). No one can serve two epistemic masters. *Philosophical Studies*, *175*(10), 2389–2398. Springer Verlag.

Gardiner, G. (2018). Legal burdens of proof and statistical evidence. In D. Coady & J. Chase (Eds.), *Routledge handbook of applied*

epistemology. Routledge.

Glass, D. H. (2002). Coherence, Explanation, and Bayesian Networks. In G. Goos, J. Hartmanis, J. van Leeuwen, M. O'Neill, R. F. E. Sutcliffe, C. Ryan, M. Eaton, et al. (Eds.), *Artificial Intelligence and Cognitive Science* (Vol. 2464, pp. 177–182). Berlin, Heidelberg: Springer Berlin Heidelberg.

Gordon, T. F., Prakken, H., & Walton, D. (2007). The Carneades model of argument and burden of proof. *Artificial Intelligence*, *171*(10-15), 875–896.

Haack, S. (2014). Legal probabilism: An epistemological dissent. In *Haack2014-HAAEMS* (pp. 47–77).

Hacking, I. (1975). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press.

Hastie, R. (2019). The case for relative plausibility theory: Promising, but insufficient. *The International Journal of Evidence & Proof*, *23*(1-2), 134–140.

Hedden, B. (2019). Hindsight bias is not a bias. *Analysis*, *79*(1), 43–52.

Hepler, A. B., Dawid, A. P., & Leucari, V. (2007). Object-oriented graphical representations of complex patterns of evidence. *Law, Probability and Risk*, *6*(1-4), 275–293.

Ho, H. L. (2008). *A philosophy of evidence law: Justice in the search for truth*. Oxford University Press.

Ho, H. L. (2019). How plausible is the relative plausibility theory of proof? *The International Journal of Evidence & Proof*, *23*(1-2), 191–197.

Horowitz, I. A., & Kirkpatrick, L. C. (1996). A concept in search of a definition: The effect of reasonable doubt instrcutions on certainty of guilt standards and jury verdicts. *Law and Human Behaviour*, *20*(6), 655–670.

Kadane, J. B., & Schum, D. A. (2011). *A probabilistic analysis of the sacco and vanzetti evidence*. John Wiley & Sons.

Kaye, D. H. (1979). The laws of probability and the law of the land. *The University of Chicago Law Review*, *47*(1), 34–56.

Kaye, D. H. (1986). The admissibility of "probability evidence" in criminal trials—part I. *Jurimetrics Journal*, 343–346.

Kaye, D. H. (2010). *The double helix and the law of evidence*. Harvard University Press.

Koehler, J. J. (1996). On conveying the probative value of DNA evidence: Frequencies, likelihood ratios, and error rates. *University of Colorado law Review*, *67*, 859–886.

Koscholke, J. (2016). Evaluating Test Cases for Probabilistic Measures of Coherence. *Erkenntnis*, *81*(1), 155–181.

Laplace, P. (1814). *Essai philosophique sur les probabilités*.

Laudan, L. (2006). *Truth, error, and criminal law: An essay in legal epistemology*. Cambridge University Press.

Lempert, R. O. (1977). Modeling relevance. *Michigan Law Review*, *75*, 1021–1057. JSTOR.

Lempert, R. O. (1986). The new evidence scholarship: Analysing the process of proof. *Boston University Law Review*, *66*, 439–477.

Lipton, P. (2004). *Inference to the best explanation*. Routledge/Taylor; Francis Group.

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Lucy, D. (2013). *Introduction to statistics for forensic scientists*. John Wiley & Sons.

Meijs, W., & Douven, I. (2007). On the alleged impossibility of coherence. *Synthese*, *157*(3), 347–360.

Merricks, T. (1995). Warrant entails truth. *Philosophy and Phenomenological Research*, *55*, 841–855.

Nance, D. A. (2019). The limitations of relative plausibility theory. *The International Journal of Evidence & Proof*, *23*(1-2), 154–160.

National Research Council. (1992). *DNA technology in forensic science* [NRC I]. Committee on DNA technology in Forensic Science, National Research Council.

Neil, M., Fenton, N., Lagnado, D., & Gill, R. D. (2019). Modelling competing legal arguments using bayesian model comparison and averaging. *Artificial Intelligence and Law*. Retrieved from https://doi.org/10.1007/s10506-019-09250-3

Neil, M., Fenton, N., & Nielson, L. (2000). Building large-scale Bayesian Networks. *The Knowledge Engineering Review*, *15*(3), 257–284.

Nesson, C. R. (1979). Reasonable doubt and permissive inferences: The value of complexity. *Harvard Law Review*, *92*(6), 1187–1225.

Newman, J. O. (1993). Beyon "reasonable doub". *New York University Law Review*, *68*(5), 979–1002.

Olsson, E. J. (2001). Why Coherence Is Not Truth-Conducive. *Analysis*, *61*(3), 236–241.

Pardo, M. S. (2018). Safety vs. Sensitivity: Possible worlds and the law of evidence. *Legal Theory*, *24*(1), 50–75.

Pardo, M. S., & Allen, R. J. (2008). Judicial proof and the best explanation. *Law and Philosophy*, *27*(3), 223–268.

Pennington, N., & Hastie, R. (1991a). A cognitive theory of juror decision making: The story model. *Cardozo Law Review*, *13*, 519–557.

Pennington, N., & Hastie, R. (1991b). A cognitive theory of juror decision making: The story model. *Cardozo Law Review*, *13*, 519–557. HeinOnline.

Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of personality and social psychology*, *62*(2), 189–204. American Psychological Association.

Pettigrew, R. (2011). Epistemic utility arguments for probabilism. In *Stanford encyclopedia of philosophy*.

Posner, R. (1973). *The economic analysis of law*. Brown & Company.

Redmayne, M. (2008). Exploring the proof paradoxes. *Legal Theory*, *14*(4), 281–309. Cambridge University Press.

Robertson, B., & Vignaux, G. A. (1995). DNA evidence: Wrong answers or wrong questions? *Genetica*, *96*, 145–152.

Ross, L. (2020). Rehabilitating statistical evidence. *Philosophy and Phenomenological Research*.

Schippers, M., & Koscholke, J. (2019). A General Framework for Probabilistic Measures of Coherence. *Studia Logica*.

Schwartz, D. S., & Sober, E. (2019). What is relative plausibility? *The International Journal of Evidence & Proof*, *23*(1-2), 198–204.

Schweizer, M. (2013). The Law Doesn't Say Much About Base Rates. *SSRN Electronic Journal*.

Shen, Q., Keppens, J., Aitken, C., Schafer, B., & Lee, M. (2007). A scenario-driven decision support system for serious crime investigation. *Law, Probability and Risk*, *5*(2), 87–117.

Shogenji, T. (1999). Is Coherence Truth Conducive? *Analysis*, *59*(4), 338–345.

Shogenji, T. (2001). Reply to akiba on the probabilistic measure of coherence. *Analysis*, *61*(2), 147–150. Oxford University Press (OUP). Retrieved from https://doi.org/10.1093/analys/61.2.147

Shogenji, T. (2006). Why does coherence appear truth-conducive? *Synthese*, *157*(3), 361–372. Springer Science; Business Media LLC. Retrieved from https://doi.org/10.1007/s11229-006-9062-8

Siebel, M. (2004). On Fitelson's measure of coherence. *Analysis*, *64*, 189–190.

Siebel, M. (2006). Against probabilistic measures of coherence. In *Coherence, truth and testimony* (pp. 43–68). Springer.

Smith, M. (2017). When does evidence suffice for conviction? *Mind*.

Swinburne, R. (2001). *Epistemic justification*. Oxford University Press.

Taroni, F., Biedermann, A., Bozza, S., Garbolino, P., & Aitken, C. (2014). *Bayesian networks for probabilistic inference and decision analysis in forensic science* (2nd ed.). John Wiley & Sons.

Thomson, J. J. (1986). Liability and individualized evidence. *Law and Contemporary Problems*, *49*(3), 199–219.

Tribe, L. H. (1971). Trial by mathematics: Precision and ritual in the legal process. *Harvard Law Review*, *84*(6), 1329–1393.

Underwood, B. D. (1977). The thumb on the scale of justice: Burdens of persuasion in criminal cases. *Yale Law Journal*, *86(7)*, 1299–1348.

Urbach, P., & Howson, C. (1993). *Scientific reasoning: The bayesian approach*. Open Court.

Urbaniak, R. (2018a). Narration in judiciary fact-finding: A probabilistic explication. *Artificial Intelligence and Law*, 1–32.

Urbaniak, R. (2018b). Narration in judiciary fact-finding: A probabilistic explication. *Artificial Intelligence and Law*, 1–32.

Urbaniak, R., Kowalewska, A., Janda, P., & Dziurosz-Serafinowicz, P. (2020). Decision-theoretic and risk-based approaches to naked statistical evidence: Some consequences and challenges. *Law, Probability and Risk*.

van Eemeren, F., & Verheij, B. (2017). Argumentation theory in formal and computational perspective. *IFCoLog Journal of Logics and Their Applications*, *4*(8), 2099–2181.

Verheij, B. (2014). To catch a thief with and without numbers: Arguments, scenarios and probabilities in evidential reasoning. *Law, Probability and Risk*, *13*(3-4), 307–325. Citeseer.

Verheij, B. (2017). Proof with and without probabilities. Correct evidential reasoning with presumptive arguments, coherent hypotheses and degrees of uncertainty. *Artificial Intelligence and Law*, 1–28. Springer.

Vlek, C. (2016). *When stories and numbers meet in court: Constructing and explaining bayesian networks for criminal cases with scenarios*. Rijksuniversiteit Groningen.

Vlek, C., Prakken, H., Renooij, S., & Verheij, B. (2013). Modeling crime scenarios in a bayesian network. In *Proceedings of the fourteenth international conference on artificial intelligence and law* (pp. 150–159). ACM.

Vlek, C., Prakken, H., Renooij, S., & Verheij, B. (2014). Building bayesian networks for legal evidence with narratives: A case study evaluation. *Artificial Intelligence and Law*, *22*, 375–421. Springer.

Vlek, C. S., Prakken, H., Renooij, S., & Bart Verheij. (2015). Representing the quality of crime scenarios in a bayesian network. In A. Rotolo (Ed.), *Legal knowledge and information systems* (pp. 133–140). IOS Press.

Wagenaar, W., Van Koppen, P., & Crombag, H. (1993). *Anchored narratives: The psychology of criminal evidence.* St Martin's Press.

Walen, A. (2015). Proof beyond a reasonable doubt: A balanced retributive account. *Louisiana Law Review*, *76*(2), 355–446.

Walton, D. N. (2002). *Legal argumentation and evidence*. Penn State University Press.

Wells, G. (1992). Naked statistical evidence of liability: Is subjective probability enough? *Journal of Personality and Social Psychology*, *62*(5), 739–752. American Psychological Association.

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

Wixted, J. T., & Wells, G. L. (2017). The Relationship Between Eyewitness Confidence and Identification Accuracy: A New Synthesis. *Psychological Science in the Public Interest*, *18*(1), 10–65.