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, probabilistic explication of and theorizing about evidence evaluation and legal decision standards, remain 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.

say sth about replicability crisis in forensic sciences at some point?

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 is the most developed one but LP is still underdeveloped —to a large extent this is so in light of various lines of criticism developed by the representatives of the other strains.

The **goal of this project** is to contribute to the **development of legal probabilism** by formulating its variant that accomodates important **insights provided by its critics**. A crucial point of criticism is that the fact-finding process should be conceptualized as **a competition of narrations**. I plan to develop methods that allow the probabilist to take this perspective, and explain how such methods allow the legal probabilist to address various objections present in the literature. The key idea is that once **narrations are represented as bayesian networks, various criteria on, features of and operations on narrations can be explicated in terms of corresponding properties of and operations on bayesian networks**. Further, the hypothesis is that such an improved framework will facilitate addressing key objections raised against LP.

The conceptual developments will be accompanied by technical accounts. **R** code capturing the technical features developed will be made available to the reader. Thus, the output will be a **unifying extended probabilistic model embracing key aspects of the narrative and argumentative approaches, susceptible to AI implementation.** The methods employed include: Bayesian statistical methods (including Bayesian approach to higher-order probability), imprecise probabilities, and Bayesian networks.

Contents

2 Significance

(state of the art, justification for tackling a specific scientific problem, justification for the pioneering nature of the project, the impact of the project results on the development of the research field and scientific discipline);

2.1 State of the art

2.1.1 Legal probabilisim

One of the functions of the trial is to resolve disputes about questions of facts. Did the defendant rob the bank? Who is the father of the child? Did this drug cause birth defects? To answer such questions, the litigants will present evidence of different kinds: eyewitness testimonies, DNA matches, epidemiological studies, etc. The evidence presented will often be in conflict with other evidence. In a bank robbery case, for example, the prosecution may present eyewitness testimony that the defendant was seen driving a truck near the bank a few minutes after the robbery took place. The defense may respond that no traces were found at the crime scene that would match the defendant. The fact-finders, judges or lay jurors, should address these conflicts by assessing and weighing the evidence, and on that basis reach a final decision. This is a difficult task. The evidence presented at trial can be complex and open to multiple interpretations, and even when it is assessed carefully, it may still lead to an incorrect verdict. How should judges and jurors respond to this uncertainty?

From among the three perspectives mentioned in the beginning, 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.

In the Middle Ages, before the advent of probability theory, there existed an informal mathematics of legal evidence (Wigmore, 1901). Formalistic procedures fixed the number of witnesses required to establish a claim. Lawyers would list ways in which items of evidence could be added or subtracted to weaken or strengthen one's case. This formalistic system fell into disrepute as the Enlightenment principle of 'free proof' gained wide acceptance (Damaška, 1995). Concurrently, the development of probability theory brought forth a new approach to weighing evidence and making decisions under uncertainty. 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 (Daston, 1988; Franklin, 2001; Hacking, 1975). Bernoulli (1713) was one of the first to formulate probabilistic rules for combining different pieces of evidence in legal cases and assessing to what extent they supported a claim of interest. He was also one of the first to suggest that decision rules at trial could be understood as probability thresholds.

Bernoulli's prescient insights attained greater popularity in the 20th century amidst the law and economics movement (Becker, 1968; Calabresi, 1961; Posner, 1973). In a seminal article, 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 rely on probability theory, specifically 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 Skeptical voices and challenges

In response to such developments, Tribe (1971) attacked what he called 'trial by mathematics'. His critique ranged from 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 to the dehumanization of trial decisions. After

Tribe, many have criticized legal probabilism on a variety of grounds, both theoretical and practical, arguing that probabilistic models are either inadequate or unhelpful (Brilmayer, 1986; Cohen, 1986; Dant, 1988, Allen (1986); Underwood, 1977).

After the discovery of DNA fingerprinting in the eighties, many legal probabilists focused on how probability theory could be used to quantify the strength of a DNA match under various circumstances (Kaye, 1986, 2010; Koehler, 1996; National Research Council, 1992; Robertson & Vignaux, 1995).

Some legal scholars and practitioners have voiced their support for legal probabilism explicitly (Tillers & Gottfried, 2007). Yet skepticism about mathematical and quantitative models of legal evidence is still widespread among prominent legal scholars and practitioners (see, for example, Allen & Pardo, 2007). Even among legal probabilists, few would think it possible to quantify precisely the probability of someone's guilt or civil liability. In response, probabilistists such as Taroni, Biedermann, Bozza, Garbolino, & Aitken (2014) suggest that the probabilistic formalism is still useful as an aid to structure and guide one's inferences under uncertainty, rather than a way to reach precise numerical assessments.

Conceptually, the probabilistic approach togehter with decision-theoretic considerations, can be used to theorize about the standard of proof and its properties. But for this project to be successfull, a proper probabilist explication of such a standard needs to be agreed upon. 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.

Perhaps the most difficult conceptual challenge to such probabilistic explications—at least, one that has galvanized philosophical attention in recent years—comes from the paradoxes of legal proof or puzzles of naked statistical evidence. In a number of seminal papers, 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.

Another problem with the proposal 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 $\Psi(H,E)$ just the threshold criterion, requiring that $\mathsf{P}(H|E)$ be sufficiently high, Ψ in general fails to satisfy this equivalence.

Arguably, these scenarios underscore a theoretical difficulty with probabilistic accounts of legal standards of proof. Many articles have been written on the topic, initially by legal scholars. In the last decade, philosophers have also joined the debate—for critical surveys see Redmayne (2008), Gardiner (2018) and Pardo (2019). Crucially, even fairly recent proposals to mend the situation (Dawid, 1987, Cheng (2012), Kaplow (2014)) on the part of the legal probabilist have failed (Urbaniak, 2019 contains a detailed analysis).

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 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 narrative approach

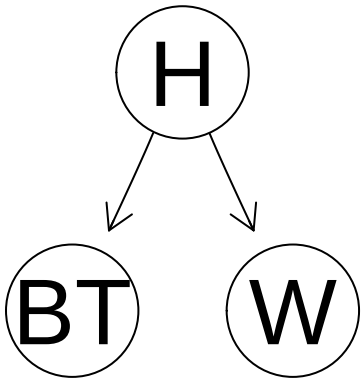
More recently, alternative frameworks for modeling evidential reasoning and decision-making at trial have been proposed. They are based on inference to the best explanation (Allen, 2010; Pardo & Allen, 2008), narratives and stories (Allen, 1986, 2010; Allen & Leiter, 2001; Clermont, 2015; Pardo, 2018; 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, 2018; Verheij, 2014).

Another important point of criticism of LP 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). 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 legal probabilism 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.

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 et al., 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 parentless nodes such as H, the tables specify the prior probabilities of all their possible states. Assuming H stands for a binary random variable, with two possible states, the prior probabilities could be:

	Prior
H=murder	.01
H=no.murder	.99

The .01 figure for H=murder rests on the assumption that, absent any incriminating evidence, the defendant is unlikely to be guilty. For children nodes, the tables specify their conditional probability given combinations of their parents' states. If the variables are binary, an assignment of values for them could be:

	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

According to the tables above, even if the defendant is not the culprit, the eyewitness testimony would still incriminate him with probability of .4, while the blood evidence with probability equal to only .063. The blood type frequency estimate is realistic (Lucy, 2013), and so are the conditional probabilities for the eyewitness identification. As expected, eyewitness testimony is assumed to be less trustworthy than blood match evidence (Urbaniak, Kowalewska, Janda, & Dziurosz-Serafinowicz, 2020; but for complications about assessing eyewitness testimony, see Wixted & Wells, 2017).

The three probability tables above are all that is needed to define the probability distribution. The tables do not specify probabilistic dependencies between nodes that are not in a relation of child/parent, such as BT and W. Since there is no arrow between them, nodes BT and W are assumed to be independent conditional on H, that is, $\mathsf{P}(()W|H) = \mathsf{P}(()W|H \wedge BT)$. This fact represents, as part of the structure of the network, the independence between eyewitness testimony and blood evidence. A generalization of this fact is the so-called Markov condition. 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.

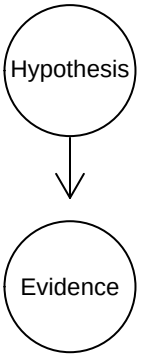
Already at this level of complexity, calculations by hand become cumbersome. In contrast, software for Bayesian networks (see, for example, the **R** package **bnlearn** developed by Marco Scutari and described in Scutari & Denis, 2015) will easily give the following results:

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

Perhaps surprisingly the posterior probability of H is about .22 even when both pieces of evidence are incriminating (BT=match, W=seen).

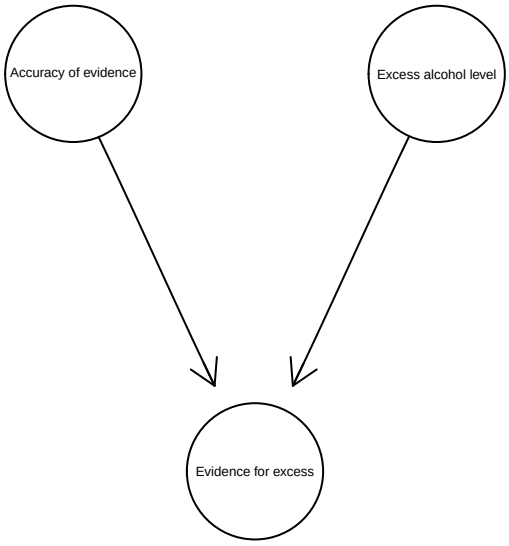
Simple graphical patterns (called *idioms*) often recur while modeling the relationships between evidence and hypotheses. Complex graphical models can be created by combining these basic patters in a modular way. Discussion of general methods for Bayesian network constructions can be found in (Neil, Fenton, & Nielson,

2000), (Hepler, Dawid, & Leucari, 2007) and general idioms are discussed in (Fenton, Neil, & Lagnado, 2013). As an example, consider the *evidence idiom* is the most basic graphical representation of the relation between a hypothesis and a piece of evidence:



This directed graph suggests that the direction of influence—which could, but need not, be interpreted as causal influence—goes from hypothesis to evidence (though the probabilistic dependence goes both ways). The hypothesis node and the evidence node can be binary variables, such as "The defendant was the source of the crime scene traces" (hypothesis) and "The defendant genetically matches the crime traces" (evidence). But the variables need not be binary. The hypothesis node might take values from the range of $1 - 40$, say the distance in meters from which the gun was shot, and the evidence node might be a continuous variable representing the density of gun shot residues (Taroni et al., 2014).

As an example of a more complex idiom, called the *evidence accuracy idiom*, consists of two arrows going into the evidence node [Bovens & Hartmann (2004); friedman1974]. One incoming arrow comes from the hypothesis node and the other from the accuracy node. This idiom can be used to model, say, an alcohol test:



The directions of the arrows indicate that the accuracy of the evidence (accuracy node) and the alcohol level (hypothesis node) influence the outcome of the test (evidence node). The graphical model represents different sources of uncertainty. The uncertainty associated with the sensitivity and specificity of the test—that is, the probability that the tests reports excessive alcohol level when the level is excessive (sensitivity) and the probability that the test reports normal alcohol level when the level is normal (specificity)—is captured by the arrow going from the hypothesis node (Excess alcohol level) to the evidence node (Evidence for excess). Other sources of uncertainty comprise the possibility that the police officer lied about the test report or the possibility that the driver took medications which then affected the alcohol level. These possibilities can be taken into consideration by adding an accuracy node (or multiple accuracy nodes, if each factor is kept separate from the other).

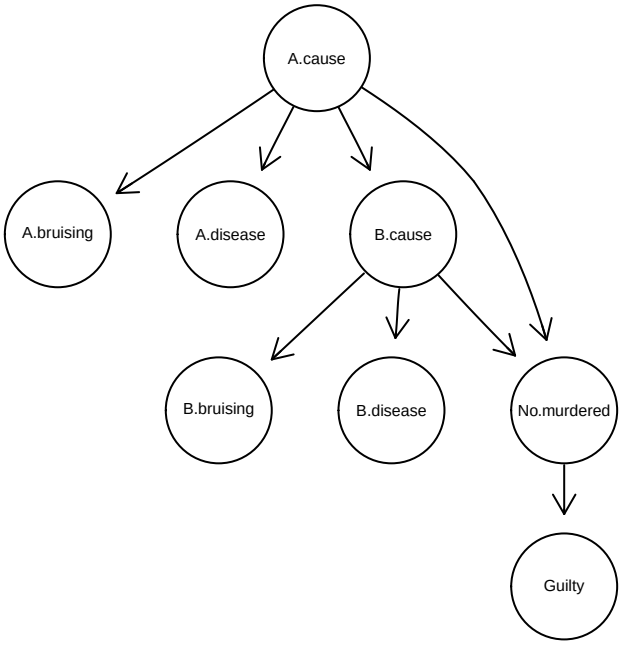
The key poin there is that large steps have been made towards the development of BN-related tools for evidence evaluation. However, so far, most of them have to do with presentation and evaluation of various pieces of evidence, not with the development of a more general model to facilitate more general probabilistic reflection on legal decision standards.

2.2 Pioneering nature of the project

2.2.1 Points of disagreement

For the reasons already mentioned, Bayesian Networks and probabilistic methods will be in the focus of this project. It is quite clear that BNs are useful tool when it comes to piecemeal modeling and evaluation of scientific evidence in court. The question is, whether they can be useful for modeling whole cases and casting light on both the conceptual challenges that we already mention and for incorporating the points made by the representatives of other strains of research, most crucially, the NPAS.

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).

According to the calculation by **?**, 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 follows:

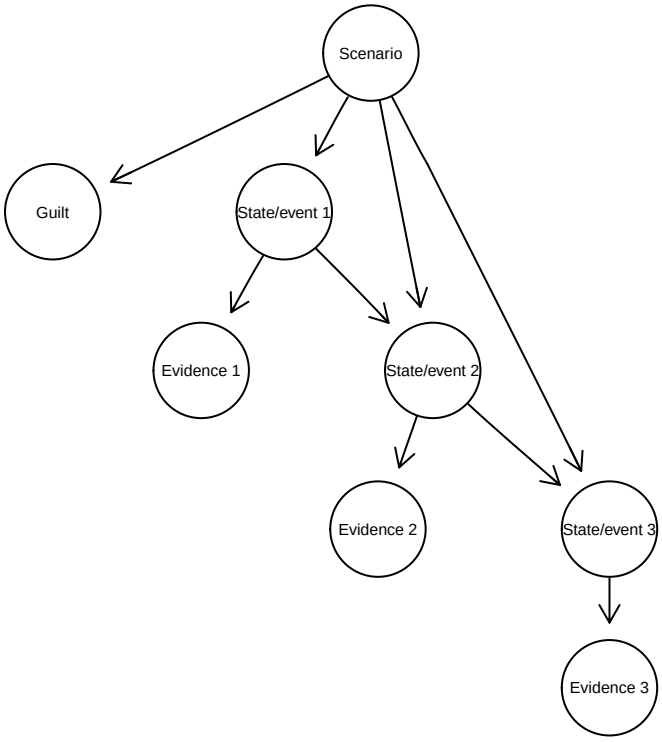
Evidence (cumulative)	P(()Clark guilty)
A bruising	.2887
A no signs of disease	.3093
B bruising	.6913
B no signs of disease	.7019

The incriminating evidence, combined, brings the probability of guilt from .0789 to .7019. This is a significant increase, but not quite enough for a conviction. If one wishes to perform sensitivity analysis by modifying some of the probabilities, this can be easily done. During the appeal trial, new evidence was discovered, in particular, evidence that son A was affected by a disease. Once this evidence is taken into account, the probability of guilt drops to .00459 (and if signs of disease were also present on B, the guilt probability would drop even further to .0009). For a general discussion on how to elicit probabilities, see (Renooij, 2001) and (Gaag, Renooij, Witteman, Aleman, & Taal, 2013).

On the other hand, the literature contains examples 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. Lacave & Díez (2002) elaborate on how BNs can be used to construct explanations.

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. The graphical model would look like this:



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, & Bart Verheij, 2015; Vlek, Prakken, Renooij, & Verheij, 2013; Vlek et al., 2014), 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. Here are the key reasons why the existing stage of development of reflection on the topic 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, by fiat 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. On one hand, 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 that it should assume. Even a single side might consider different scenarios with different probabilities, say, when there is some uncertaintly inolved in the probability assignment itself.

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, 2018): the paper only sketches some theoretical moves in a second-order language and makes no connection to BNs and so it "fails to offer a convincing and operational means to structure and compare competing narratives." 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:

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

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. Here are the key problems with this approach:

E. The assumption of equal priors is highly debatable. While 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). A criticism of the underlying reasoning of the ENFSI has has been published (Biedermann, Taroni, & Garbolino, 2007). The question remains, if the criticism is correct (as it seems to be) and the ENFSI reasoning is faulty, what should the proper application of informational entropy in the context of BN selection and averaging should look like, given that information entropy considerations independently are in epistemologically decent standing?

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

G. 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 (Dietrich & List, 2017a; see the surveys in F. & List, 2016, Dietrich & List (2017b)). One problem is that the method satsifies the unanimity assumption: whether 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. Here is a nice example. It turns out you can't at the same time hold the following: (Gallow, 2018)

$$\mathsf{P}(A = B) < 1 \tag{3}$$
$$\mathsf{P}(r|A = a) = a \tag{4}$$
$$\mathsf{P}(r|B = b) = b \tag{5}$$
$$\forall a, b\, \mathsf{P}(r|A = a, B = b) = \alpha a + \beta b \tag{6}$$

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.

2.2.2 Strategy and novelty

Representation. I will use BNs taken separately without scenario nodes to represent 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.

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 machine learning methods are not available. For instance, one cannot evaluate models in terms of their performance with respect to the data. Whether 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, no epistemological or decision-related progress might be gained based on averaging the prosecutor's and the defendant's stories. I propose ensemble methods should be deployed for multiple narration variants available from one side (as in when 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 which disagree about liability.

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 (as previously developed) to constitute priors. 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 this gives you a range of possible outcomes.

Selection criteria. While threshold- or likelihood-ratio-based selection criteria for models are unlikely to succeed, I am convinced the criteria formulated in philosophical terms in (Di Bello, 2013) and in higher-order terms in (Urbaniak, 2018) can be formulated in terms of properties of BNs. 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

(general work plan, specific research goals, results of preliminary research, risk analysis);

4 Methodology

(underlying scientific methodology, methods, techniques and research tools, methods of results analysis, equipment and devices to be used in research);

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.

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.

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.

Cheng, E. (2012). Reconceptualizing the burden of proof. *Yale LJ*, *122*, 1254. HeinOnline.

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.

Damaška, M. R. (1995). Free proof and its detractors. *The American Journal of Comparative Law*, *43*(3), 343–357.

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. (1987). The difficulty about conjunction. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *36*(2/3), 91–92.

Dawid, A. P., & Mortera, J. (2018). Graphical models for forensic analysis. In *Handbook of graphical models*

(pp. 491–514). CRC Press.

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

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

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

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

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.

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

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

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.

Gaag, L. C. van der, Renooij, S., Witteman, C. L. M., Aleman, B. M. P., & Taal, B. G. (2013). How to elicit many probabilities.

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.

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.

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

Kaplow, L. (2014). Likelihood ratio tests and legal decision rules. *American Law and Economics Review*,

16(1), 1–39. Oxford University Press.

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.

Lacave, C., & Díez, F. J. (2002). A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, *17*(2), 107–127.

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.

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

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

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

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. (2019). The paradoxes of legal proof: A critical guide. *Boston University Law Review*, *99*(1), 233–290.

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.

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.

Renooij, S. (2001). Probability elicitation for belief networks: Issues to consider. *The Knowledge Engineering Review*, *16*(3), 255–269.

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

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

Scutari, M., & Denis, J.-B. (2015). *Bayesian networks in **R***. CRC Press.

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*.

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.

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*(2), 135–157.

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.

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

Urbaniak, R. (2019). Probabilistic legal decision standards still fail. *Journal of Applied Logics*, *6*(5).

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. 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.

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.

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

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.

Wigmore, J. H. (1901). Required numbers of witnesses; a brief history of the numerical system in england. *Harvard Law Review*, *15*(2), 83–108.

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.