Chapter 1: Against Legal Probabilism

Table of contents

We present the theory of legal probabilism, discuss several objections against it and outline a number of responses available to the legal probabilist.

1 Preliminaries: The burden of proof and probability thresholds

Witnesses are called to testify in court about questions relevant to the defendant's civil or criminal liability. They can be lay people testifying about what they saw or heard. They can be experts testifying about results of laboratory testing or general scientific knowledge. As they testify, they are examined and cross-examined by the lawyers of the two parties. The rules of evidence and trial procedure frame how evidence is presented and place restrictions on certain forms of information, for example, hearsay evidence is often considered inadmissible.

Within these legal constraints, the purpose of the examination and cross-examination of witnesses is to ascertain whether the defendant engaged in behavior or committed acts that are prohibited by the applicable law. To put it somewhat crudely, the question to be answered is, did the defendant do it or not? Only if the overall evidence is strong enough to establish that the defendant did it, the defendant should be found liable.

The evidence is strong enough when it meets the governing burden of proof. This burden is different in civil or criminal cases. In civil cases, the burden of proof is 'preponderance of the evidence' (or 'balance of probabilities'); in criminal cases, the burden of proof is 'proof beyond a reasonable doubt'. The latter is meant to be more stringent than the former. These distinctions apply to countries in the common law tradition, but the concept of burden of proof seems nearly universal.

According to legal probabilism, the burden of proof is a probability threshold applied to the probability of liability. To put it more precisely, this is the probability, based on the evidence presented in court, that the defendant committed the unlawful acts or engaged in the unlawful behavior they are accused of. So, according to legal probabilism, if the evidence-based probability of liability is sufficiently high, the decision should be against the defendant, and otherwise it should favor the defendant.

Three questions naturally arise at this point. How stringent should the threshold be? How is 'probability' understood? How is the probability of liability arrived at? Consider each in turn.

1.1 First question: How stringent should the threshold be?

The stringency of the threshold—51 percent, 99 percent, or what?—is a function of maximizing expected benefits and minimizing expected costs. Think of the costs as externalities associated with false positive and false negative decisions. The benefits, instead, flow from true positive and true negative decisions. Typically, since the cost of a false positive—judging the defendant liable when they are not—is greater in criminal than civil trials, the probability threshold is set higher in criminal trials. This agrees with the fact that 'proof beyond a reasonable doubt' is more stringent than 'preponderance of the evidence'.

1.2 Second question: How is 'probability' understood

How is 'probability' to be understood in speaking of the probability that the defendant did this and that? It cannot be a long-run frequency because the acts the defendant allegedly committed are not repeatable events. It cannot be an objective chance either because the defendant either committed those acts or did not. They lie somewhere in the past. Either they occurred or did not. So the probability of liability must reflect the extent to which the evidence presented in court supports the claim that the defendant committed the unlawful acts they are accused of. The probability of liability must be evidence-relative.

1.3 Third question: How is the probability of liability arrived at?

The metaphor of a scale can be helpful. Evidence may tip the scale in one direction or the other. Evidence can point against the defendant, making it more probable that the accusation is true. Or it can point in favor of the defendant, making it less probable that the accusation is true. The balance of the scale, based on the total evidence, is the probability of liability.

Of course, this is just a metaphor. The toolkit of the legal probabilist includes more sophisticated machinery, in particular, Bayes' theorem, likelihood ratios and Bayesian networks. We will discuss them in detail in due course.

1.4 More in detail: The burden of proof in civil cases

The probabilistic picture of the burden of proof is not uncontroversial, as we shall see. But it has some plausibility, especially in civil trials. So it is worth exploring it more closely.

In civil trials, the defendant's liability is established by the balance of probabilities—the burden of proof governing civil trials—provided the defendant's liability, based on the evidence presented, is greater than .5. Or at least, this is the interpretation offered by the legal probabilist. If L stands for 'the defendant is liable' and E stans for the total evidence presented via examination and cross-examination, the burden of proof is formulated as follows:

> find against the defendant if $P(L|E) > .5$

The word 'liability' is somewhat unspecific. The defendant is accused of having committed actions or behaviors that, according to the applicable law, count as impermissible. For example, driving and drinking alcohol, committed in close temporal succession one after the other, would make one liable for driving under the influence. To represent liability at this more fine-grained level, let H_A denote the theory or hypothesis against the defendant, the accusation theory, and $\neg H_A$ its negation. The accusation theory should have some degree of specificity, for example, it should say when, where and how the defendant drove under the influence. How specific the accusation theory should be is a question we will investigate later.

So, if E is the total evidence presented in court and H_A is the accusation theory, the burden of proof in civil cases can be formulated as follows:

> find against the defendant if $P(H_A|E) > .5$ or if $P(H_A|E) > P(\neg H_A|E)$

The two formulations are equivalent because, if $P(H_A|E) > .5$, then $P(\neg H_A|E) < .5$ and thus also $P(H_A|E) > P(\neg H_A|E)$. The converse also holds. Since the probability of a proposition and its negation must add up to one, if $P(H_A|E) > P(\neg H_A|E)$, then $P(H_A|E) > .5$.

1.5 Three assumptions of legal probabilism

So, legal probabilism is committed to at least one of the following three tenets:

- the probability of liability can be assessed with some degree of precision
- the probability of liability is a good, theoretically adequate measure of the uncertainty about the disputed factual issue
- aiming to assess the probability of liability is the correct way to aggregate and weigh the evidence presented in court

We will show that any of these tenets can be challenged.

2 Challenge I: Where do the numbers come from?

2.1 The challenge in a nutshell

The metaphor of a scale tilting on one side or the other depending on the evidence presented is good only up to a point. How do we move past the metaphor?

There is broad agreement that the underlying inference engine is Bayes' theorem, but numbers must be provided to arrive at the probabilities of liability, formally modeled as the conditional probability $P(H_A|E)$. The problem is that it is difficult to assign such numbers. Here legal probabilism faces the first major challenge: *Where do the numbers come from?*

The probability of liability, assessed based on the evidence presented, is usually referred to as the posterior probability of liability. It is determined starting from a prior or initial value that is assessed before considering the evidence. The prior probability is equated to $P(L)$, while the posterior probability, given evidence E, is equated to $P(L|E)$. The relation between prior and posterior is set by the formula of Bayes' theorem:

$$P(L|E) = \frac{P(E|L)}{P(E)}P(L) = \frac{P(E|L)}{P(L)P(E|L) + P(\neg L)P(E|\neg L)}P(L)$$

The generic L could be replaced by the more specific accusation theory H_A:

$$P(H_A|E) = \frac{P(E|H_A)}{P(E)}P(H_A) = \frac{P(E|H_A)}{P(H_A)P(E|H_A) + P(\neg H_A)P(E|\neg H_A)}P(H_A)$$

The probability $P(H_A|E)$ is a function of the prior probability $P(H_A)$. It is also a function of the likelihood probabilities $P(E|H_A)$ and $P(E|\neg H_A)$. So numbers must be assigned to these probabilities to quantify $P(H_A|E)$. But it is hard to find the relevant data, and often educated guesses are the best that can done.

This challenge can be broken down into two parts: the problem of priors and that of likelihoods.

2.2 Subchallenge I: The problem of priors

The prior probability $P(L)$ must be set somewhere, but where? Should the prior be $1/n$, where n is the number of individuals who could have committed the unlawful acts in question? Setting $P(L) = 1/n$ makes sense in a criminal case in which the identity of the perpetrator is disputed. Absent information for distinguishing n possible perpetrators— before considering the trial evidence E—it is natural to set the prior probability $P(L)$ to $1/n$ since any of them could be the perpetrator. But $1/n$ does not make sense in other contexts, criminal or civil, in which the identity of the person who committed the acts isn't disputed. What is disputed,

rather, is how the acts exactly unfolded and so whether they amount to illegal conduct after all. In such cases, the prior could be l/s, where s is the number of possible ways the events could have unfolded and l is the number of legally prohibited ways, where of course $l < s$. But while it is clear how to count n, possible suspects, it is less clear how to count s or l possible ways the events could have unfolded.

2.3 Subchallenge I: The problem of likelihoods

Even if the problem of priors can be addressed, the task of assessing the likelihood probabilities $P(E|L)$ and $P(E|\neg L)$ is daunting, especially because the statements L and E are complex propositions. How is one supposed to find the right numbers?

2.4 Common responses

2.4.1 Break down and localize

It is helpful to break down L into smaller level statements, say, whether the defendant visited the crime scene or left a blood stain at the scene. It is also helpful to break down E, which refers to the total body of evidence presented at trial, into smaller components: fingerprint evidence, witness testimonies, genetic matches, expert reports, etc. Once the statements are broken down this way, assigning probabilities to them becomes more manageable.

For example, let M stand for 'the defendant's genetic profile matches the crime traces' and S stand for 'the defendant is the source of the crime trace'. The original formula reduces to the more manageable:

$$P(S|M) = \frac{P(M|S)}{P(M)}P(S) = \frac{P(M|S)}{P(C)P(M|S) + P(\neg C)P(M|\neg S)}P(S),$$

where the general level statement L is replaced by C and the overall evidence E by the match evidence M. As we will see later, $P(M|S)$ can be set to one and $P(M|\neg S)$ to the genotype probability, the expected frequency of finding the matching genotype in a reference population. The details do not matter now. The point is that these numbers can be assigned in a defensible manner. Finally, after setting $P(S) = 1/n$, where n is the number of possible contributors, the posterior probability $P(S|M)$ is obtained by easy calculations.

The strategy of focusing on smaller-level statements makes it possible to assign the numbers we need. But it also pushes the problem elsewhere. Granted, probabilities can be assigned to smaller-level statements such as M and S, but what about general-level statements such as L? The objective of a trial is not just to ascertain whether the defendant is a source of the traces found at the scene, but whether the defendant is ultimately liable.

Legal probabilists could take the hit and retreat. Focus on those domains, propositions or forms of evidence for which the required numbers are available, for example, match genetic evidence or other forms of scientific evidence. Legal probabilism need not aspire to model the evidence of an entire legal case, but only evidence amenable to probabilistic quantification.

2.4.1.1 Open questions and complications

The localization strategy is the most defensible, but limited in scope. Great progress has been made in the quantification of certain forms of scientific evidence, such as DNA matches. Experts often testify using probabilities. Hardly anyone would object that the probabilities associated with DNA matches or other forms of scientific evidence are of some use.

But larger questions would remain. How should different forms of evidence, some more easily quantifiable with probabilities than others, be aggregated? How should an entire legal case, not just small-level propositions, be modeled using probabilities? The localization strategy leaves these questions unaddressed.

2.4.2 Likelihood ratios

Forget about setting the priors for L, and only focus on the ratio of the other two probabilities, $\frac{P(E|L)}{P(E|\neg L)}$, often called the likelihood ratio. This ratio measures how strongly the evidence supports the liability claim L. The greater the ratio (for values above one), the stronger the support E lends in favor of L. So, what numbers should be assigned to $P(E|L)$ and $P(E|\neg L)$, sometimes called likelihood probabilities? As we've seen, the answer is far from clear. But, instead of assigning probabilities between 0 and 1, to both numerator and denominator, we could assign a ratio, without specifying the full probabilities. Call this the **ratio approach**.

2.4.2.1 Open questions and complications

Still, it is doubtful that ratios alone will be enough to model an entire legal case. They might serve to model uncertainty for small-level propositions, but not much beyond that.

2.4.3 Sensitivity analysis, intervals and imprecise probabilities

Another common response to the problem of setting priors is to run a **sensitivity analysis**: try out different values of the prior probability of L and see how they impact the posterior probability. If the posterior probability varies widely depending on the priors, the evidence E is weak; if it remains stable, the evidence is strong. The behavior of the posterior probability in light of setting different priors is a function of setting the other probabilities involved, specifically $P(E|L)$ and $P(E|\neg L)$. But, as seen just now, setting these other probabilities isn't an easy task either.

Sensitivity analysis can be further generalized beyond prior probabilities. Instead of precise numbers, whenever necessary, rely on **intervals** or ranges of probabilities. This is known as imprecise legal probabilism, a generalization of sensitivity analysis: test how the probability of the ultimate proposition L varies in light of ranges of values for other probabilities, including but not limited to prior probabilities.

2.4.3.1 Open questions and complications

Note that the approach based on sensitivity analysis or intervals no longer equates the burden of proof with a simple probability threshold. The stability of the posterior probability, in light of variations in the priors or other probabilities, has become an additional factor to consider.

Another problem with sensitivity analysis or the interval approach is that any probability in the interval has the same weight as any other, so extreme values are going to play too strong a role. If the interval is sufficiently large, nearly any probability value will count. More generally, imprecise legal probabilism will suffer from problems similar to imprecise probabilism in formal epistemology, such as belief inertia. So the approach rests on a shaky theoretical foundation.

2.4.4 Collect data or use expert judgment elicitation

If the numbers are missing, we do not currently have a good way to quantify uncertainty. This isn't a problem for legal probabilism; it is a problem for any procedure that attempts to ascertain disputed matters of fact. Legal probabilism has the merit of telling us what is missing. If the numbers we need are missing, that is a good reason to figure out what they are by collecting relevant data.

Return to our simple criminal case and the Bayesian network $W \leftarrow L \rightarrow S \rightarrow M$. First, we need the conditional probabilities $P(S|L)$ and $P(S|\neg L)$, focusing on the nodes in the subgraph $L \rightarrow S$. That is, if the defendant is liable (or not liable), how probable is it that they would be leaving traces at the scene? Tracking how often perpetrators manage or not to remove traces they left at the crime scene would give a number to the conditional probability $P(S|L)$, often a number close to one. The assumption is that, if someone committed the criminal act (L), it is quite likely they visited the scene at some point and left a trace (S). The closer to one this probability, the harder it is to remove the type of trace found. If the trace is invisible, it would be hard for a perpetrator to remove it. So the probability in question will depend on the type of trace that was found.

What about $P(S|\neg L)$? That is, if the defendant is not liable, how probable is it that they would still be leaving traces at the scene? Here we would need to track how often traces that have nothing to do with the crime are left at the scene. Some traces might be nearly inconsistent with innocent or causal contact, while others much less so. This will depend on the specific trace under consideration. To fix ideas, imagine the following toy model. Suppose three types of traces could be left at the crime scene—$T1$, $T2$ and $T3$—from very common

ones to traces left almost exclusively in committing criminal acts. So, finding a common $T1$ type trace would not be strongly indicative of liability, while finding a $T3$ type trace would be strongly indicative of liability. In other words, it is less likely to find $T3$ than $T2$ or $T1$ if no crime was committed. So,$P(T3|\neg L) < P(T2|\neg L) < P(T1|\neg L)$. Conversely, it is likely to find $T3$ if a crime was committed, but it is also likely to find $T1$ since it common anyway, and less common to find $T2$. So, the data nedded would consist of: different types of traces and how common they are in association with crimes and how common they are in ordinary contexts.

Second, consider the conditional probabilities $P(W|L)$ and $P(W|\neg L)$, focusing on the nodes in the subgraph $L \rightarrow W$. That is, if the defendant is liable (or not liable), how probable is it that they would be seen running away from the scene? Here we would need to collect data about the reliability of eyewitnesses in identification tasks. If a witness claims to have seen the defendant at the crime scene, how often would the witness be wrong? The numbers needed here are rates of true positive identifications, $P(W|L)$, and rates of false positive identifications, $P(W|\neg L)$. To find out these numbers, we could run experiments—as some have done—under a variety of conditions, such as distance, lighting, stress levels, duration of exposure, cross-racial settings, etc.

As these brief remarks suggest, nothing in principle bars us from collecting relevant data and entering the required numbers in the conditional probability tables.

2.4.4.1 Open questions and complications

The open question is, how many of these relevant data have been collected and are currently available? Are there ongoing research projects to make progress in this direction? Is this a promising research direction to go?

The bigger obstacle might be feasibility. There simply are too many variables to consider to collect *all* the relevant data.

For return now to the eyewitness identification case. Suppose we run experiments about false identifications, using distance, lighting, and stress level as variables. The numbers obtained from the study are then used to set the conditional probabilities $P(W|L)$ and $P(W|\neg L)$. But then, it turns out at trial that the witness had a financial stake in the dispute, and wears glasses. How should this information be added to the assessment of the probabilities? Since no other experimental study can be run during trials, the risk is that variables whose impact has not been numerically quantified in prior experiments would be neglected. This is sometimes called the problem of **soft variables**.

2.4.5 Focus on the logic of reasoning, not precise numbers

While precise numbers can be helpful, they can also be a distraction. Even without delivering precise numbers, legal probabilism is still valuable. It forces us to focus on the logic of reason-

ing under uncertainty. Probability theory imposes coherence constraints on our evidence-based beliefs about uncertain events. Similarly, legal probabilism imposes coherence constraints on evidence-based beliefs about civil and criminal liability. Legal probabilism is not primarily concerned with the task of arriving at precise probabilities. They can still be useful for illustrative purposes, but should not be taken as a basis for making decisions.

Instead of numbers, many believe that legal probabilism helps to reveal the logical structure of the reasoning, not so much, quantifying the probability of liability with exact numbers. Once the logical structure is in place, it reveals how the different sources of uncertainty add up or cancel out.

2.4.5.1 Open questions and complications

For examples of "logical structure of reasoning" see the next challenge. So the open question here is, does probability theory give us the correct and most adequate structure of reasoning under uncertainty? This is by no means obvious. See the next challenge.

3 Challenge II: Why only the probability of liability as guiding criterion?

3.1 The challenge

Even if the numerical challenge could be addressed in one of the ways outlined earlier, other difficulties exist for legal probabilism. Why ascertain the probability of liability? Is this the right metric to focus on? Intuitively, if the probability of liability were low—assuming this probability can be established—that would be a good reason not to find the defendant liable. There is little doubt about that. But this intuition only shows that a high enough probability of liability is a *necessary* condition for a finding of liability. Is it also a sufficient condition? This is less obvious. We can think of cases in which—intuitively—the probability of liability is high enough, yet the supporting evidence is weak by some other measure, and thus it is insufficient to sustain a judgment against the defendant.

3.2 Stylized case: high probability, but no liability judgment

A civil suit is brought against someone to recover a certain amount of money. This is a case of civil theft and the governing standard is preponderance or balance of probabilities. The evidence is that a rumor suggests the defendant embezzled money from the plaintiff. The accusation theory H_A is this: the defendant was working in the plaintiff's company as an administrator and in this capacity embezzled USD 50,000 of the company's funds. That the money is gone is unquestionable. It is also unquestionable that only two people had access to the money; one of them was the defendant. Without any more specific evidence about

what happened, the starting point is equipoise between H_A and its negation, where H_A is the accusation theory that the defendant embezzled the money. That is, $P(H_A) = P(\neg H_A)$. As the rumor R is added as evidence, the balance tilts towards H_A, if only so slightly. So, $P(H_A|R) > P(\neg H_A|R)$. But it would be odd—on such tenuous evidence—to conclude that the defendant is liable for embezzlement. A case like this should not even be litigated.

The rumor is insufficient evidence for several reasons. First, its reliability and trustworthiness were not tested or scrutinized. Second, the rumor did not provide a more precise account of how, when, and why the defendant embezzled the money. The rumor is certainly a relevant piece of information, but why no other evidence was presented in the case?

3.3 Resilience, specificity, completeness

Countless similar cases could be imagined.[1] The moral is that, besides high probability—or in civil cases, a greater than 50 percent probability—other dimensions (should) guide decision-making and they might not be reducible to the probability of liability. Some of these other dimensions are:

- How much should we trust the evidence presented? Was the evidence tested and scrutinized? Did it stand up to scrutiny? (**adversarial resiliency**)
- How good (specific, coherent, plausible, explanatory powerful) is the accusation theory presented? (**specificity, plausibility and coherence**)
- Is any other evidence missing? Is the evidence presented representative of both sides or was the evidence collected in a biased or skewed manner? (**evidential completeness**)

3.4 Variation on stylized case

Consider a variation of the case. Instead of a rumor, the key evidence consists in a recording of a phone conversation. In the recording, the speaker says 'I will provide you with the banking details for trasferring USD 50,000 tomorrow'. The phone conversation took place a few days before the money disappeared from the company's accounts. So the speaker must be the one who stole the company's funds. Who is the speaker, then? A voice recognition expert testifies at trial that the voice profile in the call matches the defendant's voice profile. In addition, the expert testifies that, based on a databse of 100 voice profiles, the matching profile occurs with 20 percent frequency. Is this enough to find the defedant liable of civil theft?

The starting point is $P(H_A) = P(H_D)$ given the set up of the case (only two people had access to the company's accounts). What needs to be determined is the ratio $\frac{P(PM|H_A)}{P(VM|H_D)}$, where PM is the phone recording matching that defendant's voice profile. Let $P(PM|H_A) = 1$ and

[1] *side note*: could add the Hmong drug trafficking case from Marcello's Ethics article, though this case involves statistical evidence and this may make things needlessly confusing. Another is Posner's Wall Mart case.

$P(P(PM|H_D)) = .2$.[2] So, $\frac{P(H_A|PM)}{P(H_D|PM)} = 1/.2 = 5$, largerly above the ratio threshold of 1. So should the defendant be found liable? The matching voice recording is stronger evidence than the rumor, but should we trust the expert's assessment? Is 100 instances a sufficiently large sample? Is the 20 percent figure trustworthy?

3.5 Higher order probability

Besides high probability—or in civil cases, a greater than 50 percent probability—another dimensions (should) guide decision-making and it might not be reducible to the probability of liability:

- How certain are we about the probability of liability or the posterior ratio? (**higher-order uncertainty**)

A more sophisticated version of legal probabilism, then, should be able to do at least two things: first, formally model these additional dimension using the language of probability (or determine to what extent they fall outside the scope of probability theory and cognate theories); and second, show why relying on these additional dimensions in decision-making does foster important values, such as the accuracy and fairness of trial decisions. **NEED TO EXPLAIN SECOND POINT MORE**

3.6 Sub-challenge: Trials are adversarial

Trials are often adversarial. Evidence is examined and cross-examined. How can this adversarial process be modeled probabilistically? *The chapter on cross-examination and arguments should address this challange.*

4 Challenge II: How to aggregate evidence and evaluate hypotheses?

4.1 The challenge

Legal probabilism has been successful in modeling, using probability theory, the uncertainty for small-level propositions, say the conditional probability $P(S|M)$, where \$S stands for 'source'

[2] **!!!Possible error!!!** Note sure this number is correct. If 20 percent is the frequerncy of teh voice profile in the general population, this figure does not apply to the other person who had access to the company's accounts. This other person is not the general population. So, for all we know, $P(PM|H_D) = 1$, as well. It is correct to say, P(PM vert ¬H_A)=.2, but it does not follow that P(PM |H_D)=.2. *So the comparative approach makes some numbers we have, like randm match probabilities, unusable?*

and M for 'match.' But what about general-level propositions, sometimes also called activity-level propositions? What is the probability that the defendant visited the crime scene? Or what is the probability they committed a specific act? Factual questions are often framed at different levels of analysis, and these different levels must be integrated. The question is how. A related problem is that the evidence presented in a case often consists of several pieces dependent on one another in complex ways. Some intermediate hypotheses could play the role of evidence. So, how are multiple pieces of evidence aggregated and how are complex hypotheses built out of their simpler components? This is—in a broad and rough outline—the challenge, call it the aggregation challenge.

A simple way in which this challenge can be formulated is this (inspired by the literature on the conjunction paradox; see later chapters). Suppose two propositions must be established in a case to prove liability for the defendant, A and B. Suppose $P(A|E_A) = p_A$ and and $P(B|E_B) = p_B$. Then, what is the aggregate probability $P(A \wedge B|E_A \wedge E_B)$? The answer is not to multiply p_A and p_B, but depends on the relationship between A and B, as well as E_A and E_B. The challenge, then, is to offer a systematic method to analyze these dependencies and offer a recipe for the aggregation of evidence and hypotheses.

4.2 A common response: Bayesian networks

Legal probabilists often rely on **Bayesian networks** to map out complex cases involving multiple propositions and multiple pieces of evidence. These networks serve to draw the connections between smaller-level propositions, such as S and M, and general-level ones, such as L. We will see how they work and how they are built in later chapters. A rough sketch will suffice for now.

Consider a stylized criminal case in which the items of evidence and disputed propositions are graphically represented as follows: $W \leftarrow L \rightarrow S \rightarrow M$, **DRAW BAYES NET HERE** where the letters L, S and M are interpreted as before. In addition, let W stand for an incriminating eyewitness testimony, say the testimony that they saw the defendant run away from the scene.

This network represents a case in which match evidence (M) supports the claim that the defendant left traces at the scene (S). In turn, this latter claim supports the ultimate, general-level claim L that the defendant is liable. The witness testimony (W) supports L directly. Given this setup, the probability of L given both M and W is what we are interested in, $P(L|M\&W)$.

The first major problem we encounter here is related to the earlier Challenge I. Even such a simple Bayesian network will need several conditional probabilities to get the calculations going. Besides $P(M|S)$ and $P(M|\neg S)$, it will need the conditional probabilities $P(S|L)$ and $P(S|\neg L)$. That is, if the defendant is liable (or not liable), how probable is it that they would be leaving traces at the scene? The network will also need the conditional probabilities $P(W|L)$ and $P(W|\neg L)$. That is, if the defendant is liable (or not liable), how probable is it

that they would be seen running away from the scene? These probabilities must be entered in the probability tables associated with the network. Presumably, $P(S|L)$ must be greater than $P(S|\neg L)$ and $P(W|L)$ greater than $P(W|\neg G)$. But besides these inequalities, what else?

Bayesian networks do not solve the problem of how to assign the numbers. If anything, they make the problem more apparent. It is difficult to find all the numbers required by the probability tables of a Bayesian network even in simple networks, with just a few nodes making up the network. When Bayesian networks consist of several propositions and items of evidence, the problem is even starker. So, then, the required numbers are often inserted as educated guesses because the probability tables cannot be left blank.

But let's now set aside the question of numbers. Focus on the logic or the structure of reasoning itself. Are probability theory and Bayesian networks well equipped to model the logic of evidence aggregation and hypothesis evaluation in light of the evidence? There are competing theories: argumentation theory; coherence and narratives; and relative plausibility and ruling out alternatives.

4.3 Subchallenge II: Modeling argument patterns

Bayesian networks are certainly a significant innovation and they should be used in aggregating evidence and hypotheses. They offer a rigorous and general method for this purpose. But besides the problem of assigning numbers, they also suffer from another major shortcoming: how do familiar argument patterns—that we find in argumentation theory—map onto Bayesian networks? This is far from straightforward. Here are some examples:

- A chain of uncertain inferences from A to B to C to D etc. will tend to accumulate uncertainty and become progressively weaker.
- If two or more pieces of independent evidence E1, E2, etc. support the same hypothesis H, the overall support for H should be greater than the support by E1, E1, etc. independently.
- If two or more pieces of evidence support conflicting hypotheses, and they have equal strength, their overall support cancels out.
- Attacking evidence can be rebutting (when it supports a conflicting hypothesis) or undercutting (when it undercuts the evidential support of another piece of evidence for a hypothesis) and in both cases it weakens evidential support but in different ways.
- When else?

But, hopefully, Bayesian networks will not simply validate familiar argument patterns but also offer further nuance and insight. Bayesian networks could also show that some reasoning patterns are invalid, or they could validate new reasoning patterns that informally are hard to see.

4.3.1 Example: Linear chain of uncertain inferences

Suppose the following inference structure applies to a case: $L \rightarrow S \rightarrow M$, where is the ultimate proposition and M is known, match evidence. So, the inference from M to the intermediate step S suffers from some uncertainty and the inference from M to L suffers some additional uncertainty. So, overall, the inference from M to L suffers from the aggregate uncertainty from M to S to L. Is there a qualitative way to model this aggregation of uncertainty using Bayesian networks?

This pattern can then be generalized along a linear chain of inferences that count several intermediate steps. We start with S_1, then go along a chain of inferences from S_k to the next step S_{k+1} and end up with the conclusion C. Each step is subject to some uncertainty. Knowing the uncertainties associated with each step, what is the overall uncertainty associated with the inference from S_1 to C?

In the argumentation framework, each reasoning step could be defeated. If no reasoning step is defeated, then the entire chain of inferences stands. If any one of the reasoning steps is defeated, the entire chain breaks down. This is a simple approach, but it is binary and does not allow for modeling the strength of each reasoning step.

How to model this using Bayesian networks? Suppose we know the uncertainty of the inference from A to B, namely in terms of $P(B|A)$ and we also know the uncertainty of the inference from B to C, namely in terms of $P(C|B)$. Then, what is the aggregate uncertainty of the inference from A to C via B? Is uncertainty aggregation compositional? In other words, let $P(B|A) = u_{AB}$ and $P(C|B) = u_{BC}$. What is the overall uncertainty $P(C|A) = u_{AC}$? By total probability,

$$P(C|A) = P(C|A \wedge B)P(B|A) + P(C|A \wedge \neg B)P(\neg B|A).$$

If we assume a Bayesian network structure $A \rightarrow B \rightarrow C$ (though this assumption must be justified), this simplifies to:

$$P(C|A) = P(C|B)P(B|A) + P(C|\neg B)P(\neg B|A) = u_{BC}u_{AB} + (1 - u_{AB})P(C|\neg B).$$

This is intuitive. The inference from A to C runs through two distinct paths: the path ABC, where each step in the path holds, or the path $A\overline{B}C$ in which the intermediate step B does not hold, but it still possible for C to arise even if B does not hold. This is an insight that we would not have obtained without using probabilistic modeling. How useful is this? Unclear. What are we to make of this? And can this insight be generalized to longer chains of inferences consisting of many more intermediate steps? What would uncertainty look like?

If we generalize this model, the addition of an extra intermediate step would double the number of possible paths. With one intermediate step (A-B-C), there are two possible paths; with two intermediate steps (A-B-C-D), there are $2 \cdot 2$ intermediate steps; with three intermediate steps (A-B-C-D-E), there are $2 \cdot 2 \cdot 2$ paths; and so on. So, in general, there are 2^k, for k intermediate steps. How does the uncertainty of the overall inference grow as the number of intermediate inference steps grows?

4.4 Subchallenge II: Modeling Coherence boost and fitting together

Besides familiar reasoning patterns, there are also controversial ones. For example, some believe that when different pieces of evidence fit together into a well-specified narrative—see Susan Haack's metaphor of the crosswords puzzle—this fact should give rise to a boost in evidential support. Is this true? How should this boost be understood exactly? What does it mean for pieces of evidence to 'fit together'?

5 Challenge III: how to aggregate qualitative/quantative evidence

6 Challenge IV: how to model an entire legal case?

7 Comparative thresholds

Part of the problem of assessing probabilities is having to sift through a large space of possibilities. The hypothesis L and its negation—or H_A and its negation—cover the entire sample space, and the negations $\neg L$ or $\neg H_A$ are hard to envision. A more manageable task is to start with a limited space of alternative hypotheses, assign priors to them and then adjust the probabilities in light of the evidence available. In the simplest case, only two competing hypotheses are compared against one another.

Which brings us to a competitor of legal probabilism, the theory of relative plausibility (Allen and Pardo). The starting point of this theory is that at trial the two parties put forward competing explanations of the evidence. Think of them as the accusation theory H_A and the defense theory H_D. Instead of assessing the probability of L or H_A in light of the evidence E, the theory of relative plausibility submits that the point of legal fact-finding is to assess the plausibility of H_A *relative to* H_D in light of evidence E. Plausibility is a multidimensional notion, comprising considerations of fit and consistency, predictive power, logical coherence, coverage of the evidence, etc. The more plausible explanation should prevail given a weighted combination of these criteria. So the judgment of liability should agree with the better explanation.

To articulate the idea of plausibility more precisely woud bring us too far afield. But one aspect of relative plausibility is clear: instead of focusing on the accusation H_A and assessing its probability given the evidence E, focus on comparing the accusation theory against its alternative H_D, not the full negation of the accusation theory. This comparative idea can be adopted by legal probabilism.

So what if $\neg H_A$ is replaced by a more specific alternative, namely H_D, the theory put foward by the defense? While H_D entails $\neg H_A$, because H_D and H_A must be incompatible, the converse does not hold. $\neg H_A$ does not entail H_D becuase H_D is just one particular way in which H_A

can fail to hold. So comparing H_A and H_D is generally computationally less burdensome than assessing H_A as such (which also requires comparing H_A and its full negation $\neg H_A$).

Bayes' theorem, in its ratio version, now becomes more manageable:

$$\frac{P(H_A|E)}{P(H_D|E)} = \frac{P(E|H_A)}{P(E|H_D)} \times \frac{P(H_A)}{P(H_D)}$$

Assessing the priors, the likelihood probabilities $P(E|H_A)$ and $P(E|H_D)$, and the posterior probabilities now is a more manageable task. Take the case of the priors. Instead of having to assess what the probability of L or H_A should be, in some absolute sense, all that is needed is to asess the probability of H_A relative to that of H_D, the ratio of the two.

Crucially, the posterior ratio $\frac{P(H_A|E)}{P(H_D|E)}$ does not deliver the probability of $P(H_A|E)$, unless H_D is the straight negation of H_A (which it often is not). So if the posterior ratio is, say 100-to-1, it does not follow that $P(H_A|E)$ equals $100/(100+1) \approx .99$.

That the posterior ratio cannot be use to deduce the posterior probability $P(H_A|E)$ should not be a problem, however. On the compartive approach, the goal should be to assess the ratio,not the posterior probability $P(H_A|E)$. So the rule of decision, the burden of proof, should also be understood in this comparative fashion (call it the **comparative formulation**):

> find against the defendant if $P(H_A|E) > P(H_D|E)$, or if $\frac{P(H_A|E)}{P(H_D|E)} > t$, where t is a threshold of interest, with $t = 1$ in civil cases and a greater value criminal cases.

In other words, to establish the defendant's civil liability by the balance of probabilities, the accusation theory H_A should be more probable than the defense theory. In criminal cases, the comparaison will be more stringent, but let us focus on civil cases for the sake of simplicity.

The condition $P(H_A|E) > P(H_D|E)$ is not equivalent to $P(H_A|E) > .5$ seen earlier for civil cases. It could be that both H_A and H_D have probability below .5, even though H_A is more probable than H_D. So, following this comparative formulation, a defendant could be found liable even though the probability of liability is below .5. This result seems counterintuitive, and perhaps it is a reason to favor the earlier, non-comparative formulation of the burden of proof. A related reason to be cautious of the comparative formulation is that the decision rule depends on the choice of H_A and H_D. It is possible that, given the same stock of evidence E, in one case the probability of H_A exceeds that of H_D, while in another case, given a different framing of the two theories, the probability of H_D exeeeds that of H_A. This result signals a worrisome level of sujbectivity in the decision rule.

So the comparative approach is promising, on the one hand, becaue it presumably makes assessing probabilities more manageable, but on the other, it suffers from the difficulties just outlined, for example, that a defendant would be liable even when the probability of liability is less than 50 percent. Such difficulties, however, are not indicative of a problem with the

comparative approach per se. Rather, they are characteristic of drawing inferences from evidence under conditions of uncertainty. For the non-comparative formulation of the burden of proof suffers from similar problems, as we shall now see.

Let's start with the observation that any probability assessment is always relative to a probability model. The latter comprises a sample space and a distribution over all possible outcomes (or events, propositions). Say I am interested in the probability that the sum of two six-sided fair dice is between 10 and 12 upon tossing the dice toghether once. To compute this probability, one should draw a sample space, assign probabilities to each event in the sample space and then assess the probability of the event of interest following the axims of probability. The resulting probability will be relative to the chosen sample space and the probability distribution over the events in the sample space. Any probability model should satisfy two constraints: logical coherence (it follows the probability axioms) and emprical plausibility (it is consistent with available data and our commonsense or scientific picture of the world). Even though the latter constraint is less easy to check than the first, it is no less important.

What counts as—intuitively—the same event H_A could be assigned different probabilities by two models, $M1$ and $M2$. So it is entirely possible that $P_{M1}(H_A) > .5 > P_{M2}(H_A)$. Depending on the probability model, the probability H_A may or may not meet the .5 threshold. On the non-comparative formulation of the burden of proof, then, a defendant could be found liable even though the probability of liability—according to some model, say $M2$—is below .5. On the comparative formulation, the problem was caused by a dependence on the choice of the competing or alternative theory H_D. Here the problem is caused by the choice of the probability model. These two problems are closey related. In fact, they may well be the same problem under different guises.

Suppose that in $M1$ the defense hypothesis H_D just is $\neg H_A$ and we have $P_{M1}(H_A) > 0.5 > P_{M1}(\neg H_A) = P_{M1}(H_D)$. But suppose, instead, that in $M2$, the defense hypothesis H_D is just one way to negate H_A and we have $P_{M2}(H_A) > P_{M2}(H_D)$, but both are below 0.5. There is a third hypothesis H_D^* which covers the remaining space of possibilities. Here we are assuming that $\Omega_{M1} \subseteq \Omega_{M2}$, namely that the space of possibilities in $M2$ includes all the events in $M1$ and more. In fact, in $M1$, $H_A \cup H_D = \Omega_{M1}$, even though, in $M2$, $H_A \cup H_D \cup H_D^* = \Omega_{M2}$. So what are we supposed to conclude? Is H_A above the 0.5 probabiliyt thrshold or not? It depends on the model. But since $M2$ seems a better model than $M1$ since its sample space is larger, then it would seem that H_A is actually below .5 and if we were to decide based on $M1$, we would inded find the defedant liable even the probability of liability is below .5. All in all, just as the comparative approach suffers from hypothesis-dependence, the non-comparative approach suffers from model-dependence. These depedencies on the choice of the model or the hypothesis might well be a manifestation of a common phenomenon.

8 Challenge II: Learning isn't (just) updating (or model-dependence)

The discussion about comparative thresholds and model- or hypothesis-dependence brings us to another objection to legal probabilism. In a slogan: learnign isn't (just) updating. Ronald Allen complains that Bayesian updating isn't an adequate model of what goes on in the courtroom when evidence is presented. The decision-makers do not start from priors over propositions and update them based on the pieces of evidence presented. What happens is more complicated and cannot be modeled by Bayesian updating alone. This is an important point. Certainly part of what goes in the court can be modeled using updating. The jurors might think that if evidence E is presented, then H_A should become less likely. When E is actually presented, they lower their confidence in H_A. This is updating. But much more seems to be going on. What else? New alternative hypotheses or explanations are discovered or brought to light. When this happen, the space of possibilities of sample space should be reconsidered. Bayesian updating now breaks down.

The challenge can be better articulated by identifying three related reasoning tasks that are performed in the courtroom: (i) assessing and interpreting the evidence; (ii) formulating hypotheses in light of the evidence; and (iii) assessing hypotheses in light of the evidence. Bayesian updating, if anything, can model the third task only. Once the content of the evidence is fixed and the hypotheses under consideratiuon are fixed, updating can do its job. In the end, it is a rather narrow process. In light of this, legal probabilists can respond to teh challange in two different ways:

- They can grant that Bayesian updating is limited to a narrow part of reasoning with evidence and hypotheses. Bayesian updating only models a form of *retrospective coherence*. Once the possible hypotheses and space of possibilities are fixed, as well as the evidence presented has been interpreted and framed, Bayesian updating checks that the probability assignments are coherent in light of the evidence so interpreted. But those probability assignments were already entered from the beginning (or from the end) once the sample space and the probability model had been defined, so that all $P(A|E)$ were defined.
- Legal probabilist can extent the framework beyond mere updating and cover tasks (i) assessing and interpreting the evidence and (ii) formulating hypotheses in light of the evidence. There is updatign within the model (standard Bayesin updating) and there is updating of the model (space of possibilities) outside the model. The latter seems an integral part of reasoning with evidence and hypotheses.

See the chapter on cross-examination and arguments should address this challange.

If we focus on task (iii) only, the question remains whether legal probabilism and the theory of relative plausibility are in fact in broad agreement or differences still exist.

9 Challenge IV: Evidence is evaluated holistically.

The chapter on story coherence should address this challange.

10 Challenge V: No evidence that probability reduces errors

It is clear that people make probabilistic mistakes in reasoning, but does this show that mistaken convictions are caused by these probabilistic mistakes? There is no evidence of that. In what way does probability actually improve the accuracy of legal decisions? *Discussion about accuracy and fairness should address this challange*

11 Structure

So we can envision four central chapters:

Chapter: Higher-order probability See existing chapter and paper on higher-order legal probabilism.

Chapter: Narratives, specificity, coherence etc. See Rafal's paper on coherence.

Chapter: Cross-examination and arguments See Marcello's paper on cross-examination and Bayesian networks, and also paper on awareness growth and Bayesian networks.

Chapter: Gaps in Evidence See existing paper on gaps in the evidence.

This more sophisticated version of legal probabilism should answer some of existing challenges to simple legal probabilism.

12 Things to add April 10 meeting

- weight of evidence?
- goals of legal probabilism, why do we want to use legal probabilism, what do we think we should improve? allows you to see errors or problema you would not otherwise see
- model uncertainty: how does it relate to awareness grwoth or model comparison?
- conjuction paradox, naked statistical evidence
- likelihood problems

- causally responsible for something, using causal models